Practice Problems for CSC 262

Anthony Almudevar

1 Probability

1. **The Inclusion-Exclusion Principle**. If events $A_1, \ldots, A_n$ are mutually exclusive, the probability of the union is
$$P(\cup_{i=1}^n A_i) = P(A_1) + \ldots + P(A_n).$$
If they are not mutually exclusive, then calculation of the probability of their union can become quite complex. For two events, we have
$$P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 A_2).$$
This can be extended to n events using the *inclusion-exclusion identity*
$$\begin{aligned} P(\cup_{i=1}^n A_i) &= \sum_i P(A_i) \\ &\quad - \sum_{i<j} P(A_i A_j) \\ &\quad + \sum_{i<j<k} P(A_i A_j A_k) \\ &\quad \vdots \\ &\quad -1^{n+1} P(A_1 A_2 \ldots A_n). \end{aligned}$$
   (a) Write explicitly the *inclusion-exclusion identity* for $n = 3$.
   (b) Suppose any integer from 1 to 100 inclusive is chosen at random with equal probability, which will be denoted N. What is the probability that N is divisible by at least one of 3, 4 or 7?

SOLUTION

(a)
$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(AB) - P(AC) - P(BC) + P(ABC).$$

(b) Let E_i be the event that N is divisible by i. So,
$$E_7 = \{N \in \{7, 14, 21, 28, 35, 42, 49, 56, 63, 70, 77, 84, 91, 98\}\},$$
giving
$$P(E_7) = \frac{|E_7|}{100} = \frac{14}{100}.$$
Also, we must have
$$E_i \cap E_j = E_{i \times j}, \quad E_i \cap E_j \cap E_k = E_{i \times j \times k}.$$

So, by the inclusion-exclusion identity we have

$$\begin{aligned} P(E_3 \cup E_4 \cup E_7) &= P(E_3) + P(E_4) + P(E_7) - P(E_3E_4) - P(E_3E_7) - P(E_4E_7) + P(E_3E_4E_7) \\ &= P(E_3) + P(E_4) + P(E_7) - P(E_{12}) - P(E_{21}) - P(E_{28}) + P(E_{84}). \\ &= \frac{1}{100} \times (33 + 25 + 14 - 8 - 4 - 3 + 1) = \frac{58}{100} \end{aligned}$$

2. A *random walk* can be described as follows. We have time points $i = 0, 1, 2, \ldots$. The random walk has value X_i at time point i, according to the following rules:

   (1) $X_0 = 0$.

   (2) At time point i, +1 or -1 is added to X_i with equal probability, resulting in $X_{i+1} = X_i - 1$ or $X_{i+1} = X_i + 1$. All increments are selected independently.

   For example, we could have $X_0 = 0$, $X_1 = 1$, $X_2 = 0$, $X_3 = -1$, $X_4 = -2$, $X_5 = -1$ and so on.

   Determine the following probabilities:

   (a) $P(X_1 = 1, X_2 = 0, X_3 = -1, X_4 = 0)$,

   (b) $P(X_1 = -1, X_2 = -2, X_3 = -3, X_4 = -2)$,

   (c) $P(X_4 = 0)$,

   (d) $P(X_i > 0 \text{ for } i = 1, 2, 3, 4)$.

SOLUTION The key here is to recognize that there are 16 possible paths from X_0 to X_4, which can be enumerated with a bit of effort:

Path Index	X_0	X_1	X_2	X_3	X_4
1	0	1	2	3	4
2	0	1	2	3	2
3	0	1	2	1	2
4	0	1	2	1	0
5	0	1	0	1	2
6	0	1	0	1	0
7	0	1	0	-1	0
8	0	1	0	-1	-2
9	0	-1	0	1	2
10	0	-1	0	1	0
11	0	-1	0	-1	0
12	0	-1	0	-1	-2
13	0	-1	-2	-1	0
14	0	-1	-2	-1	-2
15	0	-1	-2	-3	-2
16	0	-1	-2	-3	-4

Then, for parts (a)-(b), we are simply being asked for the probability of a single path, which must be 1/16. For parts (c)-(d) we enumerate the paths which are in the event:

(a) $P(X_1 = 1,\ X_2 = 0,\ X_3 = -1,\ X_4 = 0) = P(\text{Path Index} = 7) = 1/16,$

(b) $P(X_1 = -1,\ X_2 = -2,\ X_3 = -3,\ X_4 = -2) = P(\text{Path Index} = 15) = 1/16,$

(c) $P(X_4 = 0) = P(\text{Path Index} \in \{4, 6, 7, 10, 11, 13\}) = 6/16,$

(d) $P(X_i > 0 \text{ for } i = 1, 2, 3, 4) = P(\text{Path Index} \in \{1, 2, 3\}) = 3/16.$

3. Three 6-sided dice are tossed independently. Label the dice red, green and blue. Suppose we define the following events:

$$\begin{aligned} A_1 &= \{\text{red dice} = \text{green dice}\} \\ A_2 &= \{\text{red dice} = \text{blue dice}\} \\ A_3 &= \{\text{green dice} = \text{blue dice}\}. \end{aligned}$$

Calculate the probabilities:

$$\begin{aligned} P(A_i), &\quad i = 1, 2, 3; \\ P(A_i \cap A_j), &\quad i \neq j; \\ P(A_1 \cap A_2 \cap A_3). & \end{aligned}$$

Are the events A_1, A_2, A_3 independent? Are they pairwise independent?

SOLUTION There are $6 \times 6 = 36$ outcomes involving two dice, and the two dice are equal for 6 of them:

$$P(A_i) = 6/36 = 1/6$$

for $i = 1, 2, 3$. Next, note that for any $i \neq j$

$$A_i \cap A_j = \{\text{all three dice are equal}\} = A_1 \cap A_2 \cap A_3.$$

There are $6 \times 6 \times 6 = 216$ outcomes involving three dice, and the three dice are equal for 6 of them:

$$P(A_1 \cap A_2) = P(A_1 \cap A_3) = P(A_2 \cap A_3) = P(A_1 \cap A_2 \cap A_3) = 6/216 = 1/36.$$

This means

$$(1/6)^3 = P(A_1)P(A_2)P(A_3) \neq P(A_1 \cap A_2 \cap A_3) = 1/36,$$

so that A_1, A_2, A_3 *are not independent*. However, for any $i \neq j$,

$$(1/6)^2 = P(A_i)P(A_j) = P(A_i \cap A_j) = 1/36,$$

so that A_1, A_2, A_3 *are pairwise independent*.

4. A game is played in the following way. First a 6-sided dice is tossed. Suppose the dice shows N. Then a coin is tossed N times. The player wins if the coin shows the same face for each of the N tosses. What is the probability that the player wins? Use the law of total probability.

SOLUTION Let W be the event that the player wins. If $N = n$, then by independence

$$P(\text{All Heads}) = P(\text{All Tails}) = (1/2)^n,$$

so that

$$P(W \mid N = n) = (1/2)^n + (1/2)^n = (1/2)^{n-1}.$$

The sample space is partitioned by the 6 events $A_1 = \{N = 1\}, \ldots, A_6 = \{N = 6\}$. By the law of total probability

$$\begin{aligned} P(W) &= \sum_{i=1}^{6} P(W \mid N = i) P(N = i) \\ &= \sum_{i=1}^{6} (1/2)^{i-1} \times \frac{1}{6} \\ &= \frac{1}{6} \times (1 + 1/2 + 1/4 + 1/8 + 1/16 + 1/32) \\ &= \frac{1}{6} \times 2 \times \frac{63}{64} \\ &= \frac{21}{64}. \end{aligned}$$

5. This question is adapted from *Introduction to Probability Models* (10th Edition), S.M. Ross. Three prisoners, labeled A, B and C, are informed by a guard that one of them has been chosen at random to be executed the following day. Prisoner A asks the guard, privately, to name one of the other prisoners who will be released. We then have the competing claims:

   (a) The guard argues that by eliminating one prisoner from the execution pool the probability that A is executed changes from 1/3 to 1/2.

   (b) Prisoner A argues that since it is already known that at least one of prisoners B or C will be released, the probability that A is executed remains 1/3.

   Assume that if the guard names a prisoner to be released, and both B and C are to be released, the guard will name either one with equal probability. Otherwise, the guard names the only prisoner other than A being released. Define the following events.

$$\begin{aligned} E_A &= \{\text{Prisoner } A \text{ chosen for execution}\} \\ E_B &= \{\text{Prisoner } B \text{ chosen for execution}\} \\ E_C &= \{\text{Prisoner } C \text{ chosen for execution}\} \\ F_B &= \{\text{Guard informs prisoner } A \text{ that prisoner } B \text{ is being released}\} \\ F_C &= \{\text{Guard informs prisoner } A \text{ that prisoner } C \text{ is being released}\}. \end{aligned}$$

   So, the event that B is to be released is equivalent to E_B^c, and the event that A is informed that B is to be released is F_B. Calculate the following probabilities:

   (a) $P(E_B^c)$,

   (b) $P(F_B)$,

   (c) $P(E_A \mid E_B^c)$,

(d) $P(E_A \mid F_B)$.

Who is correct, the guard or prisoner A?

SOLUTION

(a) We have $P(E_B) = 1/3$, so $P(E_B^c) = 1 - P(E_B) = 1 - 1/3 = 2/3$.

(b) We must have

$$\begin{aligned} P(F_B \mid E_A) &= 1/2 \\ P(F_B \mid E_B) &= 0 \\ P(F_B \mid E_C) &= 1 \end{aligned}$$

so by the law of total probability:

$$\begin{aligned} P(F_B) &= P(F_B \mid E_A)P(E_A) + P(F_B \mid E_B)P(E_B) + P(F_B \mid E_C)P(E_C) \\ &= (1/2)(1/3) + 0(1/3) + 1(1/3) \\ &= 1/2. \end{aligned}$$

(c) Noting that $E_A \subset E_B^c$, and so $E_A E_B^c = E_A$ we have

$$P(E_A \mid E_B^c) = \frac{P(E_A E_B^c)}{P(E_B^c)} = \frac{P(E_A)}{P(E_B^c)} = \frac{1/3}{2/3} = 1/2.$$

(d)

$$P(E_A \mid F_B) = \frac{P(E_A F_B)}{P(F_B)} = \frac{P(F_B \mid E_A)P(E_A)}{P(F_B)} = \frac{(1/2)(1/3)}{1/2} = 1/3.$$

Prisoner A is correct.

6. A coin is tossed twice, independently. Define the three events:

$$\begin{aligned} A_1 &= \{ \text{ first toss is Heads } \} \\ A_2 &= \{ \text{ second toss is Heads } \} \\ A_3 &= \{ \text{ two outcomes are the same } \}. \end{aligned}$$

Prove that these events are pairwise independent but not independent.

SOLUTION The sample space is $S = \{HH, HT, TH, TT\}$. We have

$$\begin{aligned} A_1 &= \{HH, HT\} \\ A_2 &= \{HH, TH\} \\ A_3 &= \{HH, TT\}. \end{aligned}$$

We have $P(A_i) = 1/2$ for $i = 1, 2, 3$.

To prove **pairwise independence** it suffices to show that $P(A_iA_j) = P(A_i)P(A_j) = 1/4$ for each pair $i \neq j$. We have

$$\begin{aligned} P(A_1A_2) &= P(\{HH\}) = 1/4, \\ P(A_1A_3) &= P(\{HH\}) = 1/4, \\ P(A_2A_3) &= P(\{HH\}) = 1/4. \end{aligned}$$

establishing pairwise independence.

To disprove **independence** we note

$$P(A_1 \cap A_2 \cap A_3) = P(\{HH\}) = 1/4 \neq P(A_1)P(A_2)P(A_3) = 1/8.$$

7. The *Monty Hall problem* is a good example of the often counterintuitive nature of probability. It is based on the television game show *Let's Make a Deal* (starring Monty Hall). There are three doors. Behind one is a car, and behind the other two are goats. The contestant picks one door. Then, one the other doors is opened, revealing a goat (this can always be done, since there are two goats). The contestant is offered the choice of staying with the original choice, or switching to the one remaining door. The contestant wins whatever is behind the selected door. Assume the contestant makes the first choice at random, and has decided in advance whether or not to switch. Determine the probability of winning the car if the contest doesn't switch, and if the contestant does switch.

SOLUTION

We can, without loss of generality, assume that the car is behind door 1 (the argument is identical wherever the car is). Define events

$$\begin{aligned} D_i &= \{ \text{ Contestant initially picks door } i \}, \ \ i = 1, 2, 3. \\ W &= \{ \text{ Contestant wins car } \}. \end{aligned}$$

Then $P(D_i) = 1/3$.

If the **contestant doesn't switch**, we clearly have $P(W \mid D_1) = 1$ and $P(W \mid D_2) = P(W \mid D_3) = 0$. By conditioning on the initial selection (the Law of Total Probability) we have

$$P(W) = P(W \mid D_1)P(D_1) + P(W \mid D_2)P(D_2) + P(W \mid D_3)P(D_3) = 1 \times 1/3 + 0 \times 1/3 + 0 \times 1/3 = 1/3.$$

So the probability of winning the car is $1/3$ if the contestant doesn't switch.

If the **contestant does switch**, we have $P(W \mid D_1) = 0$, since the contestant necessarily switches to a goat. However, if the contestant initially picks door 2 (which has a goat) then it has to be door 3 which is opened. The contest would have to switch to door 1, thus winning the car. This means $P(W \mid D_2) = 1$, and also $P(W \mid D_3) = 1$ by an identical argement. By conditioning on the initial selection (the Law of Total Probability) we have

$$P(W) = P(W \mid D_1)P(D_1) + P(W \mid D_2)P(D_2) + P(W \mid D_3)P(D_3) = 0 \times 1/3 + 1 \times 1/3 + 1 \times 1/3 = 2/3.$$

So the probability of winning the car is 2/3 if the contestant does switch.

8. A dice game is played in the following way. A player continues to toss a dice as long as the current outcome is strictly higher than the previous outcome. The score is the number of such outcomes. For example, for the sequence 1,3,5,2 the player stops at the fourth toss, and scores $X = 3$. What is the probability that the player scores at least $X = 3$?

SOLUTION Noting that the player will always toss a dice at least twice, we define events

$$\begin{aligned} E &= \{ \text{ Player score at least } X = 3 \}, \\ A_{i,j} &= \{ \text{ First two tosses are } i, j \}. \end{aligned}$$

Consider event $A_{i,j}$. If $i \geq j$ then E cannot occur. If $i < j$, then E occurs with probability $(6 - j)/6$. Each event $A_{i,j}$ has probability $P(A_{i,j}) = 1/36$. This is expressed as conditional probabilities:

$$P(E \mid A_{i,j}) = \begin{cases} (6-j)/6 & ; \quad i < j \\ 0 & ; \quad i \geq j \end{cases} .$$

By conditioning on the events $A_{i,j}$ we have (including only those events for which $P(E \mid A_{i,j}) > 0$)

$$\begin{aligned} P(E) &= P(E \mid A_{1,5})P(A_{1,5}) + P(E \mid A_{2,5})P(A_{2,5}) + P(E \mid A_{3,5})P(A_{3,5}) + P(E \mid A_{4,5})P(A_{4,5}) \\ &\quad + P(E \mid A_{1,4})P(A_{1,4}) + P(E \mid A_{2,4})P(A_{2,4}) + P(E \mid A_{3,4})P(A_{3,4}) \\ &\quad + P(E \mid A_{1,3})P(A_{1,3}) + P(E \mid A_{2,3})P(A_{2,3}) \\ &\quad + P(E \mid A_{1,2})P(A_{1,2}) \\ &= \left(4 \times \frac{1}{6} \times \frac{1}{36}\right) + \left(3 \times \frac{2}{6} \times \frac{1}{36}\right) + \left(2 \times \frac{3}{6} \times \frac{1}{36}\right) + \left(1 \times \frac{4}{6} \times \frac{1}{36}\right) \\ &= \frac{4+6+6+4}{6 \times 36} = \frac{5}{54}. \end{aligned}$$

9. The hour hand on a 12-point clock is positioned at 12. The hand moves backwards or forwards one position with equal probability N times. All moves are independent. Determine the probability that the hand rests at 3 if:

   (a) $N = 9$,
   (b) $N = 10$,
   (c) $N = 19$.

SOLUTION An outcome consists of a sequence of N (F)orward or (B)ackwards directions, say, $BFF \ldots BFB$. The problem is best approached by defining a random outcome X, defined as

$$X = \text{The number of } F\text{'s in the sequence.}$$

This is because the problem can be resolved by knowing X. By the rule of product, each outcome has the same probability $1/2^N$. The number of sequences of length N that have exactly k F's is $\binom{N}{k}$, since we are making an unordered selection of k positions for the F's from the N available. So,

$$P(X = k) = \binom{N}{k}(1/2)^N.$$

To determine if the final position of the hour is 3 we use the rule

$$\begin{aligned} (\#F - \#B) \mod 12 &= 3, \quad \text{or equivalently} \\ (2X - N) \mod 12 &= 3. \end{aligned}$$

Note that $x = y \mod n$ if $y - x$ is divisible by n. So $-9 \mod 12 = 3$. The order of the moves does not matter.

(a) There are two ways for the hour hand to rest on position 3 if $N = 9$. Either it moves forward 6 and backwards 3 positions, or it moves backwards 9 positions. The values of X for which which this occurs are 0 and 6, so

$$P(\text{ hand rests on 3 }) = P(X = 0) + P(X = 6) = \frac{\binom{9}{0} + \binom{9}{6}}{2^9} = \frac{85}{512} \approx 0.166.$$

(b) The hour hand cannot rest on 3 if $N = 10$, so

$$P(\text{ hand rests on 3 }) = 0.$$

(c) For $N = 19$ the hour hand rests on 3 if $X = 5, 11$ or 17. This means

$$\begin{aligned} P(\text{ hand rests on 3 }) &= P(X = 5) + P(X = 11) + P(X = 17) \\ &= \frac{\binom{19}{5} + \binom{19}{11} + \binom{19}{17}}{2^{19}} \\ &= \frac{87381}{524288} \approx 0.167. \end{aligned}$$

Note that the answers to (a) and (c) are both close to, but not exactly, 1/6.

10. A standard 52 card playing deck assigns a unique combination of 13 ranks (2,3,4,5,6,7,8,9,10,J,Q,K,A) and 4 suits (Clubs, Diamonds, Hearts, Spades) to each card ($13 \times 4 = 52$). Suppose a hand of 5 cards is selected at random. Using the *rule of product* calculate the probability that the cards form the following hands:

(a) **One Pair.** Exactly two cards on one rank, the remaining cards of distinct rank.

(b) **Two Pairs.** Two distinct ranks represented by exactly two cards, the remaining card of distinct rank.

(c) **Three of a Kind.** Exactly three cards of one rank, the remaining cards of distinct rank.

(d) **Straight.** Five distinct ranks in sequence, with at least two suits represented. Note that $A, 2, 3, 4, 5$ is a sequence.

(e) **Flush.** All cards of the same suit, but not in rank sequence. Note that $A, 2, 3, 4, 5$ is a sequence.

(f) **Full House.** Two cards of one rank, three cards of a different rank.

(g) **Four of a Kind.** Exactly three cards of one rank, the remaining cards of distinct rank.

(h) **Straight Flush.** All cards of the same suit, and in rank sequence. Note that $A, 2, 3, 4, 5$ is a sequence. Excludes Royal Flush.

(i) **Royal Flush.** All cards of the same suit, in rank sequence $10, J, Q, K, A$.

Carefully list the *tasks* used in the application of the *rule of product*.

SOLUTION Recall that we can have

$$D = \binom{52}{5} = 2,598,960$$

possible hands. Use the rule of product for each problem:

(a) **One Pair**

1. Select rank for pair, $n_1 = 13$.
2. Select combination of 2 from 4 cards for pair rank, $n_2 = \binom{4}{2} = 6$.
3. Select 3 distinct ranks for remaining cards, $n_3 = \binom{12}{3} = 220$.
4. Select 1 of 4 suits for each of the remaining cards, $n_4 = 4^3 = 64$.

There are

$$N = n_1 \times n_2 \times n_3 \times n_4 = 13 \times 6 \times 220 \times 64 = 1,098,240$$

such selections, so

$$P(\text{ One Pair }) = \frac{1,098,240}{2,598,960} \approx 0.4226.$$

(b) **Two Pairs**

1. Select combination of 2 from 13 ranks for the pairs, $n_1 = \binom{13}{2} = 78$.
2. Select 2 from 4 cards for first pair rank, $n_2 = \binom{4}{2} = 6$.
3. Select 2 from 4 cards for second pair rank, $n_3 = \binom{4}{2} = 6$.
4. Select 1 of $44 = 52$ - 8 remaining cards, $n_4 = 44$.

There are

$$N = n_1 \times n_2 \times n_3 \times n_4 = 78 \times 6 \times 6 \times 44 = 123,552$$

such selections, so

$$P(\text{ Two Pairs }) = \frac{123,552}{2,598,960} \approx 0.04754.$$

(c) **Three of a Kind**

1. Select rank for three cards of common rank, $n_1 = 13$.
2. Select combination of 3 from 4 cards for common rank, $n_2 = \binom{4}{3} = 4$.
3. Select 2 distinct ranks for remaining cards, $n_3 = \binom{12}{2} = 66$.
4. Select 1 of 4 suits for each of the remaining cards, $n_3 = 4^2 = 16$.

There are

$$N = n_1 \times n_2 \times n_3 \times n_4 = 13 \times 4 \times 66 \times 16 = 54,912$$

such selections, so

$$P(\text{ Thee of a Kind }) = \frac{54,912}{2,598,960} \approx 0.021.$$

(d) **Straight**

1. Select from 10 possible sequences (low card A to 10), $n_1 = 10$.
2. Select a suit for each card, excluding selections of common suits, $n_2 = 4^5 - 4 = 1020$.

There are

$$N = n_1 \times n_2 = 10 \times 1020 = 10,200$$

such selections, so

$$P(\text{ Straight }) = \frac{10,200}{2,598,960} \approx 0.0039.$$

(e) **Flush**

1. Select 1 from 4 suits, $n_4 = 4$.
2. Select 5 from 13 ranks, excluding sequential ranks, $n_2 = \binom{13}{5} - 10 = 1277$.

There are

$$N = n_1 \times n_2 = 4 \times 1277 = 5,108.$$

such selections, so

$$P(\text{ Flush }) = \frac{5,108}{2,598,960} \approx 0.0020.$$

(f) **Full House**

1. Select 1 from 13 ranks for the *three of a kind*, $n_1 = 13$.
2. Select 3 from 4 cards for the *three of a kind*, $n_2 = \binom{4}{3} = 4$.
3. Select 1 from 12 remaining ranks for the *two of a a kind*, $n_3 = 12$.
4. Select 2 from 4 cards for the *two of a kind*, $n_4 = \binom{4}{2} = 6$.

There are

$$N = n_1 \times n_2 \times n_3 \times n_4 = 13 \times 4 \times 12 \times 6 = 3,744$$

such selections, so

$$P(\text{ Full House }) = \frac{3,744}{2,598,960} \approx 0.001441.$$

(g) **Four of a Kind**

1. Select rank for four cards of common rank, $n_1 = 13$.
2. Select remaining card, $n_2 = 48$.

There are

$$N = n_1 \times n_2 = 13 \times 48 = 624$$

such selections, so

$$P(\text{ Four of a Kind }) = \frac{624}{2,598,960} \approx 2.401 \times 10^{-4}.$$

(h) **Straight Flush**

1. Select from 9 possible sequences (low card A to 9), $n_1 = 9$.
2. Select a common suit, $n_2 = 4$.

There are

$$N = n_1 \times n_2 = 9 \times 4 = 36$$

such selections, so

$$P(\text{ Straight Flush }) = \frac{36}{2,598,960} \approx 1.385 \times 10^{-5}.$$

(h) **Royal Flush**

1. Select a common suit, $n_1 = 4$.

There are

$$N = n_1 = 4$$

such selections, so

$$P(\text{ Royal Flush }) = \frac{4}{2,598,960} \approx 1.539 \times 10^{-6}.$$

11. The letters in MISSISSIPPI are randomly permuted.

(a) What is the probability that there are no consecutive S's (for example ISMPISPSIIS)?

(b) What is the probability that the S's are consecutive (for example, IPSSSSIIMPI)?

SOLUTION The number of permutations of MISSISSIPPI is

$$D = \binom{11}{1,2,4,4} = \frac{11!}{2! \times 4! \times 4!} = 34650.$$

(a) Use the *rule of product*.

1. Permute the letters other than S, $n_1 = \binom{7}{1,2,4} = 105$.
2. Once the remaining 7 letters have been permuted, we need to select for each S, uniquely, a position before, after or in between the letters. There are 8 such positions, so $n_2 = \binom{8}{4} = 70$.

There are

$$N = n_1 \times n_2 = 105 \times 70$$

such permutations, so

$$P(\text{No consecutive S's}) = \frac{N}{D} = \frac{105 \times 70}{34650} \approx 0.2121.$$

(b) Permuting the letters of MISSISSIPPI such that the S's are consecutive is equivalent to replacing the 4 S's with 1 S, then counting the permutations. This gives

$$N = \binom{8}{1,1,2,4} = \frac{8!}{2! \times 4!} = 840,$$

so

$$P(\text{All S's consecutive}) = \frac{N}{D} = \frac{840}{34650} \approx 0.024.$$

12. A bin contains 5 white and 5 black balls. A random selection of 2 balls is made. Let X be the number of white balls among the 2 selected. Determine $P(X = k)$ for $k = 0, 1, 2$.

SOLUTION We may temporarily label the balls within each color $1, \ldots, 5$, so that they are all distinct. Then the total number of selections is

$$D = \binom{10}{2} = 45.$$

To enumerate the selections for which $X = k$ use the *rule of product*.

1. Selection combination of k from 5 white balls, $n_1 = \binom{5}{k}$.
2. Selection combination of $2 - k$ from 5 black balls, $n_2 = \binom{5}{2-k}$.

There are

$$N = n_1 \times n_2 = \binom{5}{k} \times \binom{5}{2-k}$$

such combinations. We then have the general expression:

$$P(X = k) = \frac{N}{D} = \frac{\binom{5}{k}\binom{5}{2-k}}{\binom{10}{2}}.$$

This gives

$$\begin{aligned} P(X = 0) &= \frac{\binom{5}{0}\binom{5}{2}}{\binom{10}{2}} = \frac{1 \times 10}{45} = \frac{2}{9} \\ P(X = 1) &= \frac{\binom{5}{1}\binom{5}{1}}{\binom{10}{2}} = \frac{5 \times 5}{45} = \frac{5}{9} \\ P(X = 2) &= \frac{\binom{5}{2}\binom{5}{0}}{\binom{10}{2}} = \frac{10 \times 1}{45} = \frac{2}{9}. \end{aligned}$$

13. A container contains 2 balls each of n colors (a total of $2n$ balls). The two balls of the same color are considered identical. Derive an expression for

$$\alpha_n = P(\text{All colors are adjacent in a random permutation of all balls}).$$

SOLUTION We can use the multinomial coefficient. There are n types of balls, with $n_i = 2$ of each type, $i = 1, \ldots, n$. The number of permutations is therefore

$$D = \binom{2n}{2, \ldots, 2} = \frac{(2n)!}{\prod_{i=1}^{n} 2!} = \frac{(2n)!}{2^n}.$$

The number of permutations for which all colors are adjacent is equal to the number of permutations of the n colors:

$$N = n!$$

So,

$$\alpha_n = \frac{N}{D} = \frac{2^n n!}{(2n)!}.$$

ALTERNATIVE SOLUTION We can use the *rule of product*. Temporarily label the balls in each color pair 1 and 2. Then, to construct a permutation with adjacent colors use the following *tasks*:

1. Select permutation of colors, $n_1 = n!$.
2. Select ordering of temporary labels within each color pair, $n_2 = 2^n$.

There are

$$N = n_1 \times n_2 = n! \times 2^n$$

such permutations. There are a total of

$$D = (2n)!$$

(temporarily labelled) permutations, so

$$\alpha_n = \frac{N}{D} = \frac{2^n n!}{(2n)!}.$$

2 Random Variables

1. This question continues Question 1 of Assignment 2. A bin contains m white and n black balls. A random selection of $k \leq m + n$ balls is made (this is referred to as *sampling without replacement*). Let X be the number of white balls among the k selected. This is known as a *hypergeometric random variable*, which we denote $X \sim hyper(m, n, k)$.

   (a) Using principles of combinatorics, derive a general expression for the PMF of X. Make sure to state exactly the support $\mathcal{S}_X$ of X.

   (b) Define a sequence of Bernoulli random variables $U_1, \ldots, U_k$, setting $U_i = 1$ if the ith selected ball is white. Expressing X as their sum, determine the mean and variance of X.

   (c) Suppose we make a selection of k balls in the same manner, except that the balls are replaced immediately after being selected, and may be selected again (this is referred to as *sampling with replacement*). Let Y be the total number of white balls selected. What distribution does Y have? Show that $E[X] = E[Y]$ and determine the ratio $var[X]/var[Y]$. Verify that $var[X] \leq var[Y]$ for $k \geq 1$ and $var[X] < var[Y]$ for $k > 1$.

   (d) Based on the comparisons of part (c), under what conditions can the distribution of Y be used to approximate the distribution of X?

   (e) A lake contains N fish. Suppose J fish are caught, tagged, then released. After a period of time, K fish are caught (assume these K fish are distinct). Suppose X of these have been previously tagged. Assuming both samples are random samples, we have

   $$X \sim hyper(J, N - J, K).$$

Derive an expression using X, J, K, denoted $\hat{N}$, which can used as an estimate of total population size N. Use $E[X]$ as a guide. This method is known as *mark and recapture*, and is commonly used to estimate population sizes.

(f) Since X is random, we would like to know how close $\hat{N}$ is to N. Once way to do this is to use a *confidence set* CS of *confidence level* $1-\alpha$. Set x_{obs} to be the observed value of X. Then let N^* be a possible value of N. Then

$$N^* \in CS \quad \text{if and only if} \quad P(Y \leq x_{obs}) > \alpha/2 \ \text{ and } \ P(Y \geq x_{obs}) > \alpha/2, \tag{1}$$

where

$$Y \sim hyper(J, N^* - J, K).$$

It may be shown that CS will consist of all integers between some lower and upper bounds, that is,

$$CS = \{N : N_L \leq N \leq N_U\}$$

for some N_L, N_U. Then

$$P(N \in CS) \geq 1 - \alpha,$$

so that the confidence set contains the true value of N with a probability of at least $1-\alpha$. Write an `R` function which accepts (X, J, K, α) as input, and outputs the bounds N_L, N_U for a confidence set CS for N of confidence level $1-\alpha$. To do this, set N_U, N_L to be the maximum and minimum values of N^*, respectively, that satisfy condition (1). If you use a search algorithm, a good starting point would be $\hat{N}$ (rounded off to the nearest integer), which would be within the bounds N_L, N_U. Make use of `R` function `phyper()`.

(g) Use your function to determine a confidence set for N when $X = 13$, $J = 200$, $K = 100$, $\alpha = 0.05$.

SOLUTION

(a) We may temporarily label the balls, so that they are all distinct. Then the total number of selections is

$$D = \binom{m+n}{k}.$$

To enumerate the selections for which $X = i$ use the *rule of product*.

1. Selection combination of i from m white balls, $n_1 = \binom{m}{i}$.
2. Selection combination of $k-i$ from n black balls, $n_2 = \binom{n}{k-i}$.

There are

$$N = n_1 \times n_2 = \binom{m}{i}\binom{n}{k-i}$$

such combinations. We then have the general expression for the PMF:

$$p_X(i) = P(X = i) = \frac{N}{D} = \frac{\binom{m}{i}\binom{n}{k-i}}{\binom{m+n}{k}}.$$

To derive the support $\mathcal{S}_X$ we note that the number of white and black balls selected (i and $k-i$, respectively) must satisfy the following inequalities:

$$\begin{aligned} 0 &\leq i \leq m \\ 0 &\leq k - i \leq n \text{ or } k - n \leq i \leq k, \end{aligned}$$

so that the support is given by

$$\mathcal{S}_X = \{i : \max(0, k-n) \le i \le \min(k, m)\}.$$

(b) There are m white balls from a total of $m+n$. Therefore, $P(U_i = 1) = p = m/(m+n)$, and $E[U_i] = p$ for $i = 1, \ldots, k$. Therefore

$$E[X] = \sum_{i=1}^{k} E[U_i] = kp = \frac{km}{m+n}.$$

Since $U_i \sim bern(p)$, we have variance

$$var[U_i] = \sigma_i^2 = p(1-p) = \frac{m}{m+n}\left[1 - \frac{m}{m+n}\right] = \frac{mn}{(m+n)^2}.$$

However, the random variables U_i are not independent. Note that for any $i \neq j$ the product $U_i U_j$ is also a Bernoulli random variable, with $U_i U_j = 1$ if and only if the ith and jth ball are both white. If we condition on the event $U_j = 1$, we effectively remove a white ball before making the next selection. Therefore,

$$P(U_i = 1 \mid U_j = 1) = \frac{m-1}{m+n-1}$$

so that

$$E[U_i U_j] = P(U_i = 1, U_j = 1) = P(U_i = 1 \mid U_j = 1)P(U_j = 1) = \left(\frac{m-1}{m+n-1}\right)\left(\frac{m}{m+n}\right)$$

and

$$\begin{aligned} cov[U_i U_j] &= E[U_i U_j] - E[U_i]E[U_j] \\ &= \left(\frac{m-1}{m+n-1}\right)\left(\frac{m}{m+n}\right) - \left(\frac{m}{m+n}\right)^2 \\ &= \frac{m}{m+n}\left[\frac{m-1}{m+n-1} - \frac{m}{m+n}\right] \\ &= -\frac{mn}{(m+n)^2(m+n-1)}. \end{aligned}$$

We use the expression

$$\begin{aligned} var[X] &= \sum_i \sigma_i^2 + 2\sum_{i<j} \sigma_{ij} \\ &= \frac{kmn}{(m+n)^2} - \frac{k(k-1)mn}{(m+n)^2(m+n-1)} \\ &= \frac{kmn}{(m+n)^2}\left(\frac{m+n-k}{m+n-1}\right). \end{aligned}$$

(c) Selections are now independent, with constant probability $p = m/(m+n)$ of selecting a white ball for each draw. Therefore $Y \sim bin(k, m/(m+n))$. Then

$$E[Y] = \frac{km}{m+n} = E[X],$$

and

$$var[Y] = kp(1-p) = \frac{kmn}{(m+n)^2}.$$

This means

$$\frac{var[X]}{var[Y]} = \frac{m+n-k}{m+n-1},$$

and the inequalities follow directly.

(d) The means of X and Y are equal for all parameters m, n, k. Otherwise, based on the ratio $var[X]/var[Y]$ the variances are approximately equal if k is small compared to $m+n$. In this case, the binomial distribution may be used to approximate the hypergeometric distribution.

(e) We have J 'white balls', $N-J$ 'black balls', and a selection without replacement of size K. This means

$$X \sim hyper(J, N-J, K).$$

We have

$$E[X] = \frac{KJ}{N},$$

and so

$$N \approx \frac{KJ}{X}.$$

(f) The condition (1) can be expressed in terms of the CDF F_Y:

$$P(Y \leq x_{obs}) > \alpha/2 \text{ if and only if } F_Y(x_{obs}) > \alpha/2$$

and

$$P(Y \geq x_{obs}) > \alpha/2 \text{ if and only if } F_Y(x_{obs} - 1) < 1 - \alpha/2.$$

Then the following function produces the confidence set:

```
cs.hyper = function(x,j,k,alpha) {

# To find NL:
# start at rounded estimate, decrement n until condition for
# inclusion in CS is no longer met.

n = round(k*j/x,0)
while (phyper(x,j,n-j,k) > alpha/2) {n = n-1}
NL = n+1

# To find NU:
# start at rounded estimate, increment n until condition for
# inclusion in CS is no longer met.

n = round(k*j/x,0)
while (phyper(x-1,j,n-j,k) < (1-alpha/2)) {n = n+1}
NU = n-1

cs = c(NL=NL,NU=NU)
return(cs)
}
```

(g) We have output:

```
> j = 200
> k = 100
> alpha = 0.05
```

```
> x = 13
> cs.hyper(x,j,k,alpha)
  NL   NU
 963 2778
```

2. Suppose X is a nonnegative random variable, that is, $P(X \geq 0) = 1$. Assume that X is a discrete random variable with sample space $S_X = \{0, 1, 2, \ldots\}$. Show that

$$E[X] = \sum_{n=0}^{\infty} \overline{F}_X(n).$$

Hint: Express X as a sum of Bernoulli random variable.

SOLUTION

(a) Set $U_n = 1$ if $n \leq X$ and $U_n = 0$ otherwise. Then

$$X = \sum_{i=1}^{\infty} U_i,$$

and so

$$E[X] = \sum_{n=1}^{\infty} E[U_n] = \sum_{n=1}^{\infty} P(X \geq n) = \sum_{n=0}^{\infty} P(X > n) = \sum_{n=0}^{\infty} \overline{F}_X(n).$$

3. Suppose $X_1, \ldots, X_n$ are independent random variables with a common CDF F_X.

(a) Show that the CDF of $Y = \max(X_1, \ldots, X_n)$ is given by $F_Y(t) = F_X^n(t)$.

(b) Suppose $X_1 \sim unif(0, 1)$. Derive the density function and mean of Y.

SOLUTION

(a) We have, by independence,

$$F_Y(t) = P(Y \leq t) = P\left(\cap_i \{X_i \leq t\}\right) = \prod_i P(\{X_i \leq t\}) = F_X^n(t).$$

(b) For the $unif(0, 1)$ distribution we have

$$F_X(t) = \begin{cases} 0 & ; \quad t < 0 \\ t & ; \quad t \in [0, 1) \\ 1 & ; \quad t \geq 1 \end{cases}$$

so

$$F_Y(t) = \begin{cases} 0 & ; \quad t < 0 \\ t^n & ; \quad t \in [0, 1) \\ 1 & ; \quad t \geq 1 \end{cases} .$$

The density function is the derivative of the CDF, so (apart from $t = 0$ and $t = 1$)

$$f_Y(t) = \frac{dF_Y(t)}{dt} = \begin{cases} nt^{n-1} & ; \quad t \in [0,1] \\ 0 & ; \quad ow \end{cases}.$$

and we have

$$E[Y] = \int_{t=0}^{1} tnt^{n-1}dt = \int_{t=0}^{1} nt^n dt = \frac{n}{n+1}t^{n+1}\Big|_0^1 = \frac{n}{n+1}.$$

4. A circle has radius R, circumference C and area A.

   (a) If $R \sim exp(1)$ derive the density function for C and A. Which of these has an exponential distribution?

   (b) If $R \sim unif(0,1)$ derive the density function for C and A. Which of these has a uniform distribution?

SOLUTION Either the method of Section 4.5.1 (CDF method) or Section 4.5.2 (One-to-one method) can be used. In fact, as shown in Section 4.5.2, they will be essentially the same method when increasing transformations are used. Below, we use the method of Section 4.5.2.

We have $C = 2\pi R$ and $A = \pi R^2$. Then use transformation rule

$$f_Y(y) = \left|\frac{dg^{-1}(y)}{dy}\right| f_X(g^{-1}(y)),$$

for transformation $Y = g(X)$. We have transformations

$$\begin{aligned} C &= g_C(R) = 2\pi R, \\ g_C^{-1}(c) &= c/(2\pi), \\ \frac{dg_C^{-1}(c)}{dc} &= (2\pi)^{-1}, \text{ and if} \\ A &= g_A(R) = \pi R^2, \text{ we have} \\ g_A^{-1}(a) &= \sqrt{a/\pi}, \\ \frac{dg_A^{-1}(a)}{dr} &= 2^{-1}(\pi a)^{-1/2}. \end{aligned}$$

   (a) We have support $[0,\infty)$ and density function $f_R(r) = \exp(-r)$ for R. The support of C and A is also $[0,\infty)$. Then we have densities

$$\begin{aligned} f_C(c) &= (2\pi)^{-1} f_R(g_C^{-1}(c)) = (2\pi)^{-1}\exp\left(-c/(2\pi)\right), \quad c \geq 0, \\ f_A(a) &= 2^{-1}(\pi a)^{-1/2} f_R(g_A^{-1}(a)) = 2^{-1}(\pi a)^{-1/2}\exp\left(-\sqrt{a/\pi}\right), \quad a \geq 0, \end{aligned}$$

   with $f_C(c) = 0$ and $f_A(a) = 0$ for $c, a < 0$. Note that $C \sim exp((2\pi)^{-1})$.

   (b) We have support $[0,1]$ and density function $f_R(r) = I\{r \in [0,1]\}$ for R. The support of C is now $[0,2\pi]$ and the support of A is now $[0,\pi]$. Then we have densities

$$\begin{aligned} f_C(c) &= (2\pi)^{-1} f_R(g_C^{-1}(c)) = (2\pi)^{-1} I\{c/(2\pi) \in [0,1]\} = (2\pi)^{-1} I\{c \in [0,2\pi]\} \\ f_A(a) &= 2^{-1}(\pi a)^{-1/2} f_R(g_A^{-1}(a)) = 2^{-1}(\pi a)^{-1/2} I\{a \in [0,\pi]\}. \end{aligned}$$

Note that $C \sim unif[0, 2\pi]$. These densities are defined on the entire real line, and the support is implicit in the indicator functions. We can also write:

$$f_C(c) = \begin{cases} (2\pi)^{-1} & ; \quad c \in [0, 2\pi] \\ 0 & ; \quad ow \end{cases}$$

and

$$f_A(a) = \begin{cases} 2^{-1}(\pi a)^{-1/2} & ; \quad a \in [0, \pi] \\ 0 & ; \quad ow \end{cases} .$$

5. Assume that over a 25 year period the mean height of adult males increased from 175.5 cm to 179.1 cm, with the standard deviation remaining constant at $\sigma = 5.84$. Suppose the minimum height requirement to join the police force remained unchanged at 172 cm. Assume the heights are normally distributed.

   (a) What proportion of adult males would not meet the minimum height requirement at the beginning and end of the 25 year period?

   (b) To what value should the minimum height be changed after 25 years in order to maintain the same proportion which meet the height requirement?

   (c) What proportion of adult males would not have met this updated requirement 25 years ago?

   (d) Repeat the first three questions using the same values, except that we'll assume that the standard deviation has increased from 5.84 to 10.0 over the 25 year period.

SOLUTION Set $\mu_1 = 175.5$, $\mu_2 = 179.1$, $\sigma = 5.84$, $\sigma_{new} = 10.0$. Then let $X_1 \sim N(\mu_1, \sigma^2)$, $X_2 \sim N(\mu_2, \sigma^2)$, $X_3 \sim N(\mu_3, \sigma^2_{new})$. Also set $Z \sim N(0, 1)$.

(a) We have

$$p_1 = P(X_1 \leq 172) = P\left(\frac{X_1 - \mu_1}{\sigma} \leq \frac{172 - \mu_1}{\sigma}\right) = P(Z \leq -0.5993) \approx 0.274,$$

$$p_2 = P(X_2 \leq 172) = P\left(\frac{X_2 - \mu_2}{\sigma} \leq \frac{172 - \mu_2}{\sigma}\right) = P(Z \leq -1.2158) \approx 0.112,$$

so that the respective proportions are p_1 and p_2.

(b) Let q_1 be the minimum height required after 25 years. We need the p_1 quantile. For a standard normal distribution this is

$$p_1 = P(Z \leq Z_{p_1})$$

which is solved by $Z_{p_1} = -0.5993$. Then the p_1 quantile for the $N(\mu_2, \sigma^2)$ distribution is

$$q_1 = X_{p_1} = \mu_2 + Z_{p_1}\sigma \approx 179.1 + (-0.5993) \times 5.84 = 175.6.$$

The new minimum height should be $q_1 = 175.6$.

(c) Let p_3 be the proportion not meeting the new minimum requirement q_1 25 years ago. Then

$$p_3 = P(X_1 \leq q_1) = P\left(\frac{X_1 - \mu_1}{\sigma} \leq \frac{175.6 - \mu_1}{\sigma}\right) = P(Z \leq 0.0171) \approx 0.507.$$

(d) The quantities affected by the change from $\sigma = 5.85$ to 10.0 are p_2, q_1 and p_3. The new values are

$$p_2' = P(X_3 \leq 172) = P\left(\frac{X_3 - \mu_2}{\sigma_{new}} \leq \frac{172 - \mu_2}{\sigma_{new}}\right) = P(Z \leq -0.71) \approx 0.239,$$

$$q_1' = X_{p_1} = \mu_2 + Z_{p_1}\sigma_{new} \approx 179.1 + (-0.5993) \times 10.0 = 173.11.$$

$$p_3' = P(X_1 \leq q_1') = P\left(\frac{X_1 - \mu_1}{\sigma} \leq \frac{173.11 - \mu_1}{\sigma}\right) = P(Z \leq -0.4092) \approx 0.341.$$

The following R script can be used to calculate the answers:

```
> ## (a)
>
> p1 = pnorm(172, mean=175.5, sd=5.84)
> p2 = pnorm(172, mean=179.1, sd=5.84)
> p1
[1] 0.2744814
> p2
[1] 0.1120394
>
> ## (b)
>
> q1 = qnorm(p1, mean=179.1, sd=5.84)
> q1
[1] 175.6
>
> ## (c)
>
> p3 = pnorm(q1, mean=175.5, sd=5.84)
> p3
[1] 0.5068309
>
> ####### (d)
>
> ## (a)
>
> pp1 = pnorm(172, mean=175.5, sd=5.84)
> pp2 = pnorm(172, mean=179.1, sd=10.0)
> pp1
[1] 0.2744814
> pp2
[1] 0.2388521
>
> ## (b)
>
> qq1 = qnorm(pp1, mean=179.1, sd=10.0)
> qq1
[1] 173.1068
>
> ## (c)
>
> pp3 = pnorm(qq1, mean=175.5, sd=5.84)
```

```
> pp3
[1] 0.3409814
>
```

6. In *Natural Inheritance* by Francis Galton, published in 1889, the paired heights of parents with their adult children were reported. The heights, in inches, of 928 children are summarized in the following table. Essentially, we have a histogram. For example, 165 of the 928 adult children have heights in the class interval $(64.7, 66.7]$.

   We are interested in determining whether or not a normal distribution would be appropriate for modeling these heights. We can extract from the table estimates of mean and standard deviation $\mu \approx 68.1$ and $\sigma \approx 2.60$ (by assuming that each datum is represented by the midpoint of its class interval). In principle, we could use the empirical rule to assess the normality of the data, except for the fact that we could not expect the quantities $\mu \pm K\sigma$, $K = 1, 2, 3$, to land exactly on the endpoints, which we would need in order to obtain the relevant empirical frequencies.

   Of course, we can use other quantities to achieve the same goal. For example, we can obtain directly from the table estimates of the CDF $P(X \leq x)$ for each endpoint $x = 60.7, 62.7, \ldots, 74.7$ and compare them directly to the values predicted by the normal distribution, that is, $P(Y \leq x)$ where $Y \sim N(\mu = 68.1, \sigma^2 = 2.60^2)$. Try this, by filling in the table. See Section 4.4.4 of *Biostatistics: A Methodology for the Health Sciences* [L.D. Fisher & G. van Belle] for more on this problem.

Class Interval	Freq.	Cumulative Freq.	Estimated CDF	CDF Predicted by Normal Distribution
(58.7, 60.7]	0	-	-	-
(60.7, 62.7]	12	-	-	-
(62.7, 64.7]	91	-	-	-
(64.7, 66.7]	165	-	-	-
(66.7, 68.7]	258	-	-	-
(68.7, 70.7]	266	-	-	-
(70.7, 72.7]	105	-	-	-
(72.7, 74.7]	31	-	-	-

SOLUTION

The requires numbers can be calculated with the following R script:

```
> x0 = c(58.7, 60.7,62.7,64.7,66.7,68.7,70.7,72.7)
> x = c(60.7,62.7,64.7,66.7,68.7,70.7,72.7,74.7)
> c1 = paste('(',x0,',',x,']', sep='')
> y = c(0,12,91,165,258,266,105,31)
> tab = data.frame(c1,y,cumsum(y),round(cumsum(y)/sum(y),3),
      round(pnorm(x, mean=68.1, sd=2.6),3))
> names(tab) = paste('column',1:5)
> tab
     column 1 column 2 column 3 column 4 column 5
1 (58.7,60.7]        0        0    0.000    0.002
2 (60.7,62.7]       12       12    0.013    0.019
3 (62.7,64.7]       91      103    0.111    0.095
```

```
4 (64.7,66.7]      165      268    0.289    0.295
5 (66.7,68.7]      258      526    0.567    0.591
6 (68.7,70.7]      266      792    0.853    0.841
7 (70.7,72.7]      105      897    0.967    0.962
8 (72.7,74.7]       31      928    1.000    0.994
>
```

This gives:

Class Interval	Freq.	Cumulative Freq.	Estimated CDF	CDF Predicted by Normal Distribution
(58.7,60.7]	0	0	0.00	0.00
(60.7,62.7]	12	12	0.01	0.02
(62.7,64.7]	91	103	0.11	0.10
(64.7,66.7]	165	268	0.29	0.29
(66.7,68.7]	258	526	0.57	0.59
(68.7,70.7]	266	792	0.85	0.84
(70.7,72.7]	105	897	0.97	0.96
(72.7,74.7]	31	928	1.00	0.99

The estimated cumulative frequencies are reasonably close to the frequencies predicted by the normal distribution.

7. The file `statepop.csv` (posted in the **Assignments** folder of Blackboard) is a comma delimited text file with 50 records, each consisting of a state name and the state's population (June 2014).

   (a) Read the file into the `R` environment using the `read.table()` function.

   (b) For each state, calculate the population proportion.

   (c) Rank the proportions in decreasing order, then construct a *log-log* plot, as in Figures 4.11-4.12 of the lecture notes.

   (d) Using the `lines()` function, superimpose on this plot the lines

   $$f(k) = \log(p_X(1)) - \alpha \log(k),$$

   for $\alpha = 1/3, 2/3, 1$, where k represents the rank of a frequency ($k = 1$ is the largest frequency), and $-\alpha$ is an exponent of a power law. Use the `legend()` function to label the superimposed lines. To do this, select distinct `lty` parameters for the 3 superimposed lines, then use these values in the `legend()` function.

   (e) Do state population sizes conform to a power law? Note that the power law may hold for the largest frequencies, but not the smallest.

SOLUTION

The following `R` script produces Figure 4 below:

```
> tab = read.table('statepop.csv',header=FALSE,stringsAsFactors=FALSE,sep=',')
> names(tab) = c("state","pop")
>
> freq = tab$pop/sum(tab$pop)
>
```

```
> y = log(rev(sort(freq)))
>
> par(mfrow=c(1,1),cex=1)
> n = 50
> plot(log(1:n),y,xlab = 'log rank', ylab = 'log frequency')
> alpha = 1
> lines(log(1:n),max(log(freq)) - alpha*log(1:n),lty=1)
> alpha = 2/3
> lines(log(1:n),max(log(freq)) - alpha*log(1:n),lty=2)
> alpha = 1/3
> lines(log(1:n),max(log(freq)) - alpha*log(1:n),lty=3)
> legend('bottomleft',legend=paste('alpha = ',c('1','2/3','1/3'),sep=''),lty=1:3)
```

The highest ranking frequencies appear to conform to a power law $p_X(k) \propto 1/k^{2/3}$.

8. Suppose a casino has a game in which a player bets x dollars, then with probability p wins back $2x$ dollars (for a net gain of x) and loses the original x dollars with probability $1-p$ (for a net loss of x). Usually, $p < 1/2$. If $p = 1/2$ then the game is *fair*. Probability theory states that in such a fair game, there can be no strategy that results in a positive expected gain.

   A commonly claimed counter-example to this is the following strategy. Enter the casino, then play the game, betting $x = 1$ each time, until you have a total gain of 1. For example, the following Win/Loss sequence will accomplish this: *LWLLWWW*, which has gain sequence -1,0,-1,-2,-1,0,1, taking 7 games to reach a gain of 1. The Win/Loss sequence W also achieves a gain of 1 after a single game. Probability theory also states that the probability that a gain of 1 is reached after a finite number of games is 1 (although this *doesn't* hold if $p < 1/2$).

   This seems to lead to a contradiction, since if we use this strategy, we can play once a day, and guarantee ourselves a regular income, noting that we can use any value of x we wish. Note that the case of the fair game, $p = 1/2$, is the important one, since if no winning strategy exists for this case, no winning strategy can exist when $p < 1/2$, which settles the matter.

   (a) Write an `R` program which simulates this process. Assume $p = 1/2$. For a single simulation, start at $gain = 0$, then increase or decrease $gain$ after each game by 1. This is the *random walk* introduced in Assignment 1, Question 2. The process stops when $gain = 1$. Store the number of games T needed to reach $gain = 1$. You may use the `rbinom()` function. Truncate the process at 1000 games. If $gain = 1$ has not been reached, indicate this by setting the number of games at, say, $T = 1001$.

   (b) Repeat the simulation to get 1000 replicates of T. Estimate the PMF $p_T(k) = P(T = k)$ directly from the data. Construct a *log-log* plot of $\log(p_T(k))$ against $\log(k)$. Note that the frequencies are not sorted in this case. How many times did T exceed 1000? How many times was T within 10 games, inclusive?

   (c) Using the `lines()` function, superimpose on this plot the lines

   $$f(k) = \log(p_X(1)) - \alpha \log(k),$$

   for $\alpha = 1.0, 1.25, 2.0$. Label your plot with the `legend()` function as in Question 4. If you can conclude that T conforms to a power law, what can be said about $E[T]$? The rate at which this strategy earns money is

   $$\text{gain rate} = \frac{gain}{\text{number of games played}}.$$

   At what rate does this strategy earn money?

SOLUTION

The following R script produces the plot in Figure 5.

```
> nsim = 1000
> tt = rep(NA, nsim)
> bank = rep(NA, nsim)
>
> for (i in 1:nsim) {
+
+  x = rbinom(1000,size=1,prob=1/2)
+  z = cumsum(2*x-1)
+
+  if (sum(z==1) > 0) {
+  tt[i] = min(which(z==1))
+  bank[i] = min(z[1:tt[i]])
+  }
+  else
+  {
+  tt[i] = 1001
+  bank[i] = min(z)
+  }
+ }
>
> sum(tt <= 10)
[1] 752
> sum(tt == 1001)
[1] 32
>
> xx = as.integer(names(table(tt)))
> par(mfrow=c(1,1),cex=1)
> plot(log(xx), log(table(tt)/1000),xlab = 'log T', ylab = 'log frequency')
> lines(log(xx), max(log(table(tt)/1000)) - 1*log(xx),lty=1)
> lines(log(xx), max(log(table(tt)/1000)) - 1.25*log(xx),lty=2)
> lines(log(xx), max(log(table(tt)/1000)) - 2*log(xx),lty=3)
> legend('bottomleft',legend=paste('alpha = ',c('1.0','1.25','2.0'),sep=''),lty=1:3)
>
```

In this simulation there were 752/1000 simulated values of T within 10, and 32/1000 greater than 1000. Results will vary. From Figure 5 the power law $p_X(k) \propto 1/k^\alpha$ holds approximately, with $\alpha \approx 1.25$, and more generally with $\alpha < 2$. We then have, for some constant c,

$$E[T] = \sum_{k=1}^{\infty} k \frac{c}{k^\alpha}$$

noting that the support of T is unbounded. However, $E[T] < \infty$ only if $\alpha > 2$ (compare the summation to the integral $\int_1^\infty x^{-\alpha} dx$). If $\alpha < 2$ then in our case $E[T] = \infty$. This means that although the gambler can win a gain of 1 with probability 1 in a finite amount of time, the *gain rate* is 0, since $E[T] = \infty$. If we let G_n be the total gain after the nth game (not day), we would find

$$\lim_{n\to\infty} \frac{G_n}{n} = 0.$$

As a practical matter, using this strategy, we would often find that we cannot play enough games in a single day to achieve the daily gain of 1.

3 Stochastic Processes

1. Whether or not a Markov chain is an adequate model for a given application is an important question. We'll use a Markov chain model to design a simple tic-tac-toe player. It will play both sides.

   Create in `R` the following objects:

   (a) The board will consist of a vector of length 9. An unoccupied position is set to 0, otherwise the position is occupied by player 1 or 2.

   (b) A tic-tac-toe board has 8 'rows'. The three horizontal rows are (1,2,3), (4,5,6) and (7,8,9), the three vertical rows are (1,4,7), (2,5,8) and (3,6,9). The diagonal rows are (1,5,9) and (3,5,7). Create an 8×3 table which stores these rows.

   (c) Each position on the board belongs to certain rows. For example, position 6 belongs to rows (3,6,9) and (4,5,6), and so on. Create a list of length 9, in which the ith element is a vector of indices referencing the rows to which position i belongs.

   (d) To choose a move, player 1 examines each position, and assigns each a score. If position i is occupied it is assigned score 0. Otherwise, each row containing i is looked up. The number of positions in that row occupied by 1 and 2 are stored in `n.us` and `n.them` respectively. The row is scored according to the following table:

		`n.them` = 0	1	2
`n.us` =	0	10	100	10,000
	1	1,000	1	0
	2	100,000	0	0

   Then, the score for position i is the sum of the scores of the rows containing i. For example, for the following board it is player 1's turn to move. Positions 5 and 7 are scored 0, since they are occupied. To score position 1, note that it is contained in 3 rows, (1,2,3), (1,4,7) and (1,5,9). For row (1,2,3), `n.us` = 0, `n.them` = 0, so this row contributes 10 to the score. For row (1,4,7) `n.us` = 0, `n.them` = 1, and for row (1,5,9) `n.us` = 1, `n.them` = 0, so these rows contribute 100 and 1,000, respectively. The total score for position 1 is then 1,000 + 100 + 10 = 1,110.

$$\begin{matrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 2 & 0 & 0 \end{matrix}$$

   (e) After each position's score is calculated the position with the highest score is selected. If more than one position has the maximum score, one of these is chosen at random.

   (f) Player 2 uses the same strategy, calculating `n.us` and `n.them` accordingly.

   (g) Write an R program to simulate a tic-tac-toe game with alternating players using the same strategy. The game ends after one player completely occupies any row (and therefore wins), or the board is full. Run the simulation 1,000 times, and store the frequency of outcomes (player 1 wins, player 2 wins or the games ends in a draw). What are the frequencies of each outcome?

   (h) OPTIONAL BONUS QUESTION: Which of the three outcomes can occur? Justify your answer. Symmetry plays a role here.

   SOLUTION

The following R program plays the game as defined in Assignment 5. All games in the 1,000 simulations end in draws.

```
#### create row.table: 8x3 matrix of row definitions

row.table = matrix( c(1,2,3,4,5,6,7,8,9,1,4,7,2,5,8,3,6,9,1,5,9,3,5,7), ncol=3, byrow=T)

### create row.list. The ith element is a vector of indices to all
### rows in row.table in which position i is located

row.list = vector('list',9)
row.list[[1]] = c(1,4,7)
row.list[[2]] = c(1,5)
row.list[[3]] = c(1,6,8)
row.list[[4]] = c(2,4)
row.list[[5]] = c(2,5,7,8)
row.list[[6]] = c(2,6)
row.list[[7]] = c(3,4,8)
row.list[[8]] = c(3,5)
row.list[[9]] = c(3,6,7)

### score.matrix is a 3 x 3 matrix. Element score.matrix[n.us+1, n.them+1]
### gives the score for (n.us, n.them)

score.matrix = matrix(0, nrow=3, ncol=3)
score.matrix[3,1] = 100000
score.matrix[1,3] = 10000
score.matrix[2,1] = 1000
score.matrix[1,2] = 100
score.matrix[1,1] = 10
score.matrix[2,2] = 1

### choose.move is a function which accepts the current board,
### the values us.them = c(n.us, n.them), and the objects
### row.table, row.list, score.matrix created above.
### The score is calculated for each position. The highest score is identified.
### If more than one position has the highest score, one of them is chosen at random.
### The output is a list with elements names move (the selected position),
### max.score (the maximum score),
### score.temp (the vector of scores for each position)

choose.move = function(board, us.them, row.table, row.list, score.matrix) {

     score.temp = rep(0,9)

     # calculate score for each position

     for (i in 1:9) {
          if (board[i] == 0) {
               for (j in 1:length(row.list[[i]]) ) {
                    row.temp = board[row.table[row.list[[i]][j], ]]
```

```
                    n.us = sum(row.temp==us.them[1])
                    n.them =  sum(row.temp==us.them[2])
                    score.temp[i] = score.temp[i] + score.matrix[n.us+1,n.them+1]
               }
          }
     }

     # determine maximum score

     max.score = max(score.temp)

     # identify positions with maximum score

     move.list = which(score.temp==max.score)

     if (length(move.list)==1) {

               # if highest scoring position is unique, copy onto move

               move = move.list[1]
     } else
     {

               # otherwise, select position at random from highest scoring ones

               move = sample(which(score.temp==max.score),1)
     }

     # return selected position and score

     return(list(move=move, max.score=max.score, score.temp=score.temp))

}

### simulate nsim games

nsim = 1000

### store result in sv 0=draw, 1=Player 1 wins, 2 = Player 2 wins

sv = rep(0, nsim)

for (iii in 1:nsim) {

     # create playing board as vector of length 9

     board = rep(0,9)

     # flag==1 is used to indicate end of game

     flag = 0
```

```
     while (flag == 0) {

          # Player 1 plays

          junk = choose.move(board, c(1,2), row.table, row.list, score.matrix)

          # update board

          board[junk$move] = 1

          # Player 1 wins if score is 100000. Game ends if there are no empty positions on the board

          if ( (junk$max.score >= 100000) | (sum(board==0)==0) ) {
               flag=1
               if (junk$max.score >= 100000) {sv[iii] = 1}
          }

          # Player 2 plays if flag==0

          if (flag == 0) {

               # Player 2 plays

               junk = choose.move(board, c(2,1), row.table, row.list, score.matrix)

               # update board

               board[junk$move] = 2

               # Player 2 wins if score is 100000. Game ends if there are no empty positions on the bo
               if ( (junk$max.score >= 100000) | (sum(board==0)==0) ) {
                         flag=1
                         if (junk$max.score >= 100000) {sv[iii] = 2}
               }
          }
     }
}

# Examine results

table(sv)
```

After the program is run we should see something like this:

```
> table(sv)
sv
   0
1000
```

OPTIONAL QUESTION

The table below shows the sequence of one game, including scores for each position, and the subsequent move. For the first move, note that when the board is empty, the middle position (5) is scored highest, so Player 1 always starts there.

For Player 2's first move, all four diagonal positions (1,3,7,9) are scored highest. Player 2 selects one of these at random. However, note that by symmetry there is no important difference between these moves.

For Player 1's second move, the two remaining diagonal positions which share a row with Player 2's first position are scored highest. Player 1 chooses one of these at random. Both are essentially identical, after symmetry is accounted for.

If we continue in this way for the remaining moves, we see that after symmetry is accounted for, there is really only one available move at each turn. All games must be essentially identical, and will therefore end in draws.

	Score			Board		
1	30	20	30	0	0	0
	20	40	20	0	1	0
	30	20	30	0	0	0
2	120	110	120	0	0	0
	110	0	110	0	1	0
	120	110	120	0	0	2
3	21	1010	1110	0	0	0
	1010	0	1100	0	1	0
	1110	1100	0	1	0	2
4	111	110	11010	0	0	2
	200	0	1100	0	1	0
	0	101	0	1	0	2
5	1101	1100	0	0	0	2
	2000	0	11000	0	1	1
	0	1001	0	1	0	2
6	1101	1100	0	0	0	2
	10100	0	0	2	1	1
	0	101	0	1	0	2
7	102	1100	0	0	1	2
	0	0	0	2	1	1
	0	1001	0	1	0	2
8	3	0	0	0	1	2
	0	0	0	2	1	1
	0	10001	0	1	2	2
9	3	0	0	1	1	2
	0	0	0	2	1	1
	0	0	0	1	2	2

4 Bayes Theorem and Classification

1. A test for Hepatitis-B is developed. The test is administered to a test group of 147 individuals known to have Hepatitis-B. Of this group 123 test positive. The test is also administered to a control group of 220 subjects known to be free of Hepatitis-B. Of these, 15 test positive.

   (a) Estimate the sensitivity and specificity of the test directly from the data.

(b) This test is intended to be used in clinical populations of varying infection prevalence. Use `R` to construct plots of PPV and NPV for values of prevalence ranging from 0 to 5%. Use the `type = 'l'` option of the `plot()` function.

(c) Calculate prevalence, NPV and PPV directly from the data. How do these values compare to those shown in the plots of part (b)?

SOLUTION

We can summarize the study with the following contingency table:

Table 1: Outcomes of hepatitis-B diagnostic test for Problem 8

		Hepatitis-B		
		Positive	Negative	Total
Diagnostic Test	Positive	123	15	138
	Negative	24	205	229
	Total	147	220	367

(a) We have

$$\begin{aligned} sens &= \frac{TP}{TP+FN} = \frac{123}{147} \approx 0.837 \\ spec &= \frac{TN}{TN+FP} = \frac{205}{220} \approx 0.932. \end{aligned}$$

(b) The script shown below produces the plot in Figure 3.

(c) Directly from the table we have

$$\begin{aligned} prev &= \frac{TP+FN}{N} = \frac{147}{367} \approx 0.401 \\ PPV &= \frac{TP}{TP+FP} = \frac{123}{138} \approx 0.891 \\ NPV &= \frac{TN}{TN+FN} = \frac{205}{229} \approx 0.895. \end{aligned}$$

There is a prevalence of 40.1%, which is (hopefully) much higher than any prevalence we would expect to see in any population. The PPV estimated directly from the study is much higher than a PPV we would expect to encounter in a population, while the NPV is lower.

```
> sens = 123/147
> spec = (220-15)/220
>
> prev = seq(0,0.05,by = 0.001)
>
> par(mfrow=c(2,1),cex=1.0)
>
> f0 = function(prev,sens,spec) { sens*prev/(sens*prev + (1-spec)*(1-prev)) }
> plot(prev, f0(prev,sens,spec),type='l', xlab='Prevalence',ylab='PPV')
```

```
> title('Hepatiti-B Test')
>
>
> f0 = function(prev,sens,spec) { spec*(1-prev)/(spec*(1-prev) + (1-sens)*prev) }
> plot(prev, f0(prev,sens,spec),type='l', xlab='Prevalence',ylab='NPV')
> title('Hepatiti-B Test')
```

5 Inference for Single Population Means

1. For an *iid* sample from a normal distribution we are given sample mean $\bar{X} = 20.292$, $n = 100$, standard deviation $\sigma = 0.34$. Calculate a confidence interval for population mean μ with confidence level $1 - \alpha = 0.95$.

   SOLUTION We have $\alpha = 0.05$, so we need critical value

   $$z_{\alpha/2} = z_{0.025} = 1.96,$$

   giving level $1 - \alpha$ confidence interval

   $$\begin{aligned} CI &= \bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \\ &= 20.292 \pm 1.96 \times 0.034 \\ &= 20.292 \pm 0.0666 = (20.225, 20.359). \end{aligned}$$

2. For an *iid* sample from a normal distribution we are given sample mean $\bar{X} = 179.012$, $n = 7$, standard deviation $\sigma = 19.6$. Calculate a confidence interval for population mean μ with confidence level $1 - \alpha = 0.9$.

   SOLUTION We have $\alpha = 0.1$, so we need critical value

   $$z_{\alpha/2} = z_{0.05} = 1.645,$$

   giving level $1 - \alpha$ confidence interval

   $$\begin{aligned} CI &= \bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \\ &= 179.012 \pm 1.645 \times 7.408 \\ &= 179.012 \pm 12.185 = (166.827, 191.197). \end{aligned}$$

3. For an *iid* sample from a normal distribution we are given sample mean $\bar{X} = 2.349$, $n = 23$, standard deviation $\sigma = 0.03$. Calculate a confidence interval for population mean μ with confidence level $1 - \alpha = 0.99$.

   SOLUTION We have $\alpha = 0.01$, so we need critical value

   $$z_{\alpha/2} = z_{0.005} = 2.576,$$

   giving level $1 - \alpha$ confidence interval

   $$\begin{aligned} CI &= \bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \\ &= 2.349 \pm 2.576 \times 0.00626 \\ &= 2.349 \pm 0.0161 = (2.333, 2.365). \end{aligned}$$

¿

6 Power Curves

1. Consider the t-test for two independent samples with respective sample sizes n_1, n_2, assuming equal variances $\sigma_1^2 = \sigma_2^2 = \sigma^2$ (Section 14.1.2 of lecture notes). We wish to test null hypothesis $H_o : \mu_1 \geq \mu_2$ against alternative hypothesis $H_a : \mu_1 < \mu_2$. Then H_o is rejected for large values of test statistic:

$$T_{obs} = \frac{\bar{X}_2 - \bar{X}_1}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

where S_p^2 is the pooled variance.

   (a) Suppose $\mu_2 - \mu_1 = \Delta \neq 0$. Show that T_{obs} has a non-central t-distribution with non-centrality parameter:

$$ncp = \frac{\Delta}{\sigma}\sqrt{\frac{n_1 n_2}{n_1 + n_2}},$$

   assuming that the populations are normally distributed. One consequence of this is that $(n_1 + n_2 - 2)S_p^2/\sigma^2 \sim \chi^2_{n_1+n_2-2}$, and $\bar{X}_2 - \bar{X}_1 \perp S_p^2$.

   (b) Given the form of the non-centrality parameter ncp, show that if the total sample size $N = n_1 + n_2$ is fixed, the most powerful test is obtained by the balanced design $n_1 = n_2 = N/2$, assuming N is even.

   (c) Construct power curves for the one-sided pooled variance t-test just described (similar to Figure 17.2), for $\alpha = 0.05$, with the following features:

      i. Assume a balanced design $n = n_1 = n_2$. The plot will superimpose power curves for $n = 5, 10, 15, 20$ on the same plot.

      ii. The power curve will plot power $1 - \beta$ for one-sided alternatives $\Delta = \mu_2 - \mu_1 > 0$ against Δ/σ, over the range $0 \leq \Delta/\sigma \leq 3$ (use increments of 0.1).

      iii. Label each curve appropriately using the `text()` function.

      iv. The vertical axis should be labeled $1 - \beta$ and the horizontal axis should be labeled Δ/σ, making using the `expression()` function. See `help(plotmath)` for more detail.

      v. A grid should be superimposed with grid size 0.05 for the vertical axis and 0.125 for the horizontal axis. You can use the `abline()` function. Setting option `col = 'gray'` seems to work well.

   (d) Suppose we need to determine a per-sample sample size n for a one-sided two-sample pooled variance t-test, testing null hypothesis $H_o : \mu_1 \geq \mu_2$ against alternative hypothesis $H_a : \mu_1 < \mu_2$. A power of 90% is needed for an alternative $\mu_2 - \mu_1 = 5.85$, assuming standard deviation $\sigma = 5.2$ and using $\alpha = 0.05$. Using your power curves, what value of n would you recommend (select from 5,10,15,20)?

SOLUTION

(a) The non-central t distribution is constructed from

$$T_\delta = \frac{Z + \delta}{\sqrt{W/\nu}}$$

where $Z \sim N(0,1)$, $W \sim \chi^2_\nu$, $Z \perp W$ and δ is the non-centrality parameter. Then we can write

$$\begin{aligned} T_{obs} &= \frac{\bar{X}_2 - \bar{X}_1}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \\ &= \frac{\bar{X}_2 - \bar{X}_1 - \Delta + \Delta}{S_p\sqrt{\frac{n_1 n_2}{n_1+n_2}}} \\ &= \frac{\frac{\bar{X}_2-\bar{X}_1-\Delta}{\sigma\sqrt{\frac{n_1 n_2}{n_1+n_2}}} + \frac{\Delta}{\sigma}\sqrt{\frac{n_1 n_2}{n_1+n_2}}}{S_p/\sigma} \\ &= \frac{Z + \frac{\Delta}{\sigma}\sqrt{\frac{n_1 n_2}{n_1+n_2}}}{\sqrt{W/(n_1+n_2-2)}} \end{aligned}$$

where

$$\begin{aligned} Z &= \frac{\bar{X}_2 - \bar{X}_1 - \Delta}{\sigma\sqrt{\frac{n_1 n_2}{n_1+n_2}}} \\ W &= S_p^2/\sigma^2, \end{aligned}$$

with $Z \sim N(0,1)$, $W \sim \chi^2_{n_1+n_2-2}$, $Z \perp W$ and

$$\delta = \frac{\Delta}{\sigma}\sqrt{\frac{n_1 n_2}{n_1+n_2}}.$$

(b) A larger non-centrality parameter implies greater power. If $N = n_1 + n_2$ is fixed, we can write $n_2 = N - n_1$. Then maximize

$$\frac{n_1 n_2}{n_1+n_2} = \frac{n_1(N-n_1)}{N}$$

with respect to n_1. Taking the derivative verifies that the quantity is maximized by setting $n_1 = N/2$.

(c) The following code produces the power curves in Figure 4.

```
### set up labels with mathematical typesetting

ex0 = expression(paste("One-sided two-sample t-test with sample size ",
          italic(n)," per sample and ",alpha," = 0.05",sep=''))
ex1 = expression(paste(Delta,'/',sigma))
ex2 = expression(1-beta)

### grid for horizontal axis

del = seq(0,3,by = 0.1)

### set up graphics window, draw empty plot (type='n')

par(mar=c(4,5,2,2), oma=c(4,4,4,4), cex=1,cex.axis=1.2, cex.lab=1.2, cex.main=1.2)
plot(range(del), c(0,1), xlab=ex1, ylab=ex2, type='n')
title(ex0)
```

```
### grid

for (x in seq(0,3,by=0.125)) {abline(v=x,col='gray')}
for (y in seq(0,1,by=0.05)) {abline(h=y,col='gray')}

for (n in c(5,10,15,20)) {

   # power curve for n

   nfactor = sqrt(n/2)
   alpha = 0.05
   t.crit = qt(1 - alpha,df=2*n-2)
   y = 1 - pt(t.crit,ncp=nfactor*del,df=2*n-2)
   lines(del, y, type='l')

   # label individual plots

   ex3 = bquote(italic(n) == .(n))
   text(1,y[del==1],ex3)
}
```

(d) We have

$$\frac{\Delta}{\sigma} = \frac{5.85}{5.2} = 1.125.$$

The first curve to equal or exceed $1 - \beta = 0.9$ as n increases is $n = 15$, which is the best choice.

2. (Exercise 17.1 from lecture notes). Examine the accuracy of the approximations given in (17.8) - (17.9) by constructing two-sided power curves ($\alpha = 0.05$) for alternatives $p \in (1/2, 1)$ against null hypothesis $p_0 = 1/2$ for sample sizes $n = 10, 100, 1000$. Specifically, draw a function of $1 - \beta$ against $p \in [1/2, 1)$, using for β the exact expression in (17.7) and the approximation in (17.8) on the same plot (include the point $p = p_0 = 1/2$ in your plot). Do this separately for $n = 10, 100, 1000$. Use different line types (use the option `lty`) and use the `legend()` function to label the line types. Also, draw horizontal lines at $1 - \beta = 0.025$ and $1 - \beta = 0.05$. Explain the discrepancy when $p \approx p_0$. Given that we are generally interested in evaluating type II errors not greater than $\beta = 0.2$, do these approximations seem sufficiently accurate?

SOLUTION

The following code produces the power curves in Figure 5. The required power curves are based on the following functions:

$$\begin{aligned}
\beta_U(p, p_0, n, \alpha) &= \Phi\left(\frac{z_\alpha\sqrt{p_0(1-p_0)} - \sqrt{n}(p - p_0)}{\sqrt{p(1-p)}}\right), \\
\beta_L(p, p_0, n, \alpha) &= 1 - \Phi\left(\frac{-z_\alpha\sqrt{p_0(1-p_0)} - \sqrt{n}(p - p_0)}{\sqrt{p(1-p)}}\right), \\
\beta_{two}(p, p_0, n, \alpha) &= \beta_U(p, p_0, n, \alpha/2) + \beta_L(p, p_0, n, \alpha/2) - 1, \\
\beta_{two}(p, p_0, n, \alpha) &\approx \beta_U(p \mid p_0, n, \alpha/2) \text{ for } p > p_0.
\end{aligned}$$

Here, Φ is the CDF for the standard normal distribution. The functions β_U, β_L, β_{two} are coded as functions `bu`, `bl`, `bb`.

We expect the approximate power curve to equal $1 - \beta = 0.025$ for p close to p_0, instead of the correct value of 0.05 (recall that α is the probability of rejecting the null hypothesis at $p = p_0$). However, this discrepancy is noticeable only near p_0. For values that would be of practical interest for a power calculation ($1 - \beta \geq 0.8$), the approximation is nearly exact.

```
### Functions bu, bl, bb correspond to (17.4), (17.6) and (17.7) from lecture notes.

### Type II error for upper-tailed test

bu = function(p,p0,n,alpha) {
   zc = qnorm(1-alpha)
   sd0 = sqrt(p0*(1-p0))
   sd1 = sqrt(p*(1-p))
   ans = pnorm( (zc*sd0 - sqrt(n)*(p-p0))/sd1 )
   return(ans)
}


### Type II error for lower-tailed test

bl = function(p,p0,n,alpha) {
   zc = qnorm(1-alpha)
   sd0 = sqrt(p0*(1-p0))
   sd1 = sqrt(p*(1-p))
   ans = 1 - pnorm( (-zc*sd0 - sqrt(n)*(p-p0))/sd1 )
   return(ans)
}

### Type II error for two-sided test

bb = function(p,p0,n,alpha) { bu(p,p0,n,alpha/2) + bl(p,p0,n,alpha/2) - 1}

### construct power curves

x = seq(0.5,0.99,by=0.01)

ex1 = expression(italic(p))
ex2 = expression(1-beta)

### separate plots for n - 10, 100, 1000

# exact power curve will have lty = 1 (solid line)
# approximate power curve will have lty = 2 (dashed line)

par(mfrow=c(2,2),cex=1,oma=c(1,1,1,1))

n = 10
plot(x, 1 - bb(x,0.5,n,0.05),type='l', ylim=c(0,1),xlab=ex1,ylab=ex2)
lines(x, 1 - bu(x,0.5,n,0.05/2),type='l',lty=2)
```

```
title(bquote(italic(n) == .(n)))
abline(h=0.05)
abline(h= 0.025)
legend('topleft',legend=c('Exact power','Approximate power'),lty=c(1,2))

n = 100
plot(x, 1 - bb(x,0.5,n,0.05),type='l', ylim=c(0,1),xlab=ex1,ylab=ex2)
lines(x, 1 - bu(x,0.5,n,0.05/2),type='l',lty=2)
title(bquote(italic(n) == .(n)))
abline(h=0.05)
abline(h= 0.025)


n = 1000
plot(x, 1 - bb(x,0.5,n,0.05),type='l', ylim=c(0,1),xlab=ex1,ylab=ex2)
lines(x, 1 - bu(x,0.5,n,0.05/2),type='l',lty=2)
title(bquote(italic(n) == .(n)))
abline(h=0.05)
abline(h= 0.025)
```

7 Inference for Variances

1. For an *iid* sample from a normal distribution we are given sample standard deviation $S = 13.65$, with sample size $n = 68$. Calculate a confidence interval for population standard deviation σ, using confidence level $1 - \alpha = 0.95$. Also give the level $1 - \alpha$ lower and upper confidence bounds.

   SOLUTION The level $1 - \alpha$ confidence interval for σ is given by

$$\frac{S}{\sqrt{\left(\chi^2_{n-1,\alpha/2}\right)/(n-1)}} < \sigma < \frac{S}{\sqrt{\left(\chi^2_{n-1,1-\alpha/2}\right)/(n-1)}}.$$

   We use critical values

$$\chi^2_{n-1,\alpha/2} = \chi^2_{67,0.025} = 91.519 \text{ and } \chi^2_{n-1,1-\alpha/2} = \chi^2_{67,0.975} = 46.261.$$

   The confidence interval is then given by

$$\frac{13.65}{\sqrt{91.519/67}} < \sigma < \frac{13.65}{\sqrt{46.261/67}}$$

   or equivalently, $CI = (11.679, 16.427)$.

   The level $1 - \alpha$ lower bound for σ is given by ,

$$\sigma > \frac{S}{\sqrt{\left(\chi^2_{n-1,\alpha}\right)/(n-1)}}.$$

   The appropriate critical value is $\chi^2_{n-1,\alpha} = \chi^2_{67,0.05} = 87.108$. The lower bound is then given by,

$$\sigma > \frac{13.65}{\sqrt{87.108/67}} = 11.971.$$

The level $1-\alpha$ upper bound for σ is given by ,

$$\sigma > \frac{S}{\sqrt{\left(\chi^2_{n-1,1-\alpha}\right)/(n-1)}}.$$

The appropriate critical value is $\chi^2_{n-1,1-\alpha} = \chi^2_{67,0.95} = 49.162$. The upper bound is then given by,

$$\sigma < \frac{13.65}{\sqrt{49.162/67}} = 15.935.$$

2. We are given an *iid* sample from a normal distribution

$$64.2, 26.8, 40.4, 51.2, 30.7,$$

of sample size $n = 5$. Calculate a confidence interval for population standard deviation σ, using confidence level $1-\alpha = 0.9$. Also give the level $1-\alpha$ lower and upper confidence bounds.

SOLUTION The sample standard deviation is $S = 15.302$. The level $1-\alpha$ confidence interval for σ is given by

$$\frac{S}{\sqrt{\left(\chi^2_{n-1,\alpha/2}\right)/(n-1)}} < \sigma < \frac{S}{\sqrt{\left(\chi^2_{n-1,1-\alpha/2}\right)/(n-1)}}.$$

We use critical values

$$\chi^2_{n-1,\alpha/2} = \chi^2_{4,0.05} = 9.488 \text{ and } \chi^2_{n-1,1-\alpha/2} = \chi^2_{4,0.95} = 0.711.$$

The confidence interval is then given by

$$\frac{15.302}{\sqrt{9.488/4}} < \sigma < \frac{15.302}{\sqrt{0.711/4}}$$

or equivalently, $CI = (9.936, 36.302)$.

The level $1-\alpha$ lower bound for σ is given by ,

$$\sigma > \frac{S}{\sqrt{\left(\chi^2_{n-1,\alpha}\right)/(n-1)}}.$$

The appropriate critical value is $\chi^2_{n-1,\alpha} = \chi^2_{4,0.1} = 7.779$. The lower bound is then given by,

$$\sigma > \frac{15.302}{\sqrt{9.488/4}} = 10.972.$$

The level $1-\alpha$ upper bound for σ is given by ,

$$\sigma > \frac{S}{\sqrt{\left(\chi^2_{n-1,1-\alpha}\right)/(n-1)}}.$$

The appropriate critical value is $\chi^2_{n-1,1-\alpha} = \chi^2_{4,0.9} = 1.064$. The upper bound is then given by,

$$\sigma < \frac{15.302}{\sqrt{0.711/4}} = 29.674.$$

3. A sample of size $n = 6$ from a normal distribution $N(\mu, \sigma^2)$ is collected:

```
x = c(100.583, 99.600, 100.045, 98.963, 100.313, 100.097)
```

(a) Construct a level 95% confidence interval for σ. Verify your answer using `R`.

(b) Construct a level 95% upper confidence bound for σ. Verify your answer using `R`.

(c) Suppose the data are obtained from a pilot study. The object is to estimate the sample size required to estimate μ from the same population with a margin of error of E_0, with confidence level $1 - \alpha$. Using the upper bound of the confidence interval of part (a), it is determined that a sample size of $n = 100$ is required. If the upper confidence bound of part (b) had been used instead, what would the estimated sample size have been?

SOLUTION

(a) The level $1 - \alpha$ confidence interval is given by

$$\sigma \in \left(\frac{S_n}{\sqrt{\frac{\chi^2_{\alpha/2,n-1}}{n-1}}}, \frac{S_n}{\sqrt{\frac{\chi^2_{1-\alpha/2,n-1}}{n-1}}} \right).$$

The code below calculates the quantities $n = 6$, $S = 0.5758693$, $\chi^2_{1-\alpha/2,n-1} = 0.8312116$, $\chi^2_{\alpha/2,n-1} = 12.8325020$, giving level 95% confidence interval $(0.3594623, 1.4123852)$.

```
> x = c(100.583,99.600,100.045,98.963,100.313,100.097)
> n = length(x)
>
> # critical values from chi.sq distribution with df = n-1
>
> cl = qchisq(0.025,df=n-1)
> cu = qchisq(0.975,df=n-1)
>
> # construct level 95% confidence  interval
>
> sd0 = sd(x)
> ci = sd0*sqrt((n-1)*c(1/cu,1/cl))
> c(n,sd0,cl,cu,ci)
[1]  6.0000000  0.5758693  0.8312116 12.8325020  0.3594623  1.4123852
>
```

(b) The level $1 - \alpha$ upper confidence bound is given by

$$\sigma \leq \frac{S_n}{\sqrt{\frac{\chi^2_{1-\alpha,n-1}}{n-1}}}.$$

The code below calculates $\chi^2_{1-\alpha,n-1} = 1.145476$, giving level 95% upper confidence bound 1.203139.

```
>
> # construct level 95% upper confidence bound
>
```

```
> cl = qchisq(0.05,df=n-1)
> ucb = sd0*sqrt((n-1)/cl)
> c(cl,ucb)
[1] 1.145476 1.203139
```

(c) The formula for the sample size required for a level $1-\alpha$ confidence level with margin of error E_o is

$$n = \left(z_{\alpha/2} \frac{\sigma}{E_o} \right)^2$$

To compare the sample sizes n_1, n_2 based on two alternative standard deviations σ_1, σ_2, we take the ratio

$$\frac{n_2}{n_1} = \frac{\left(z_{\alpha/2} \frac{\sigma_2}{E_o} \right)^2}{\left(z_{\alpha/2} \frac{\sigma_1}{E_o} \right)^2} = \frac{\sigma_2^2}{\sigma_1^2}.$$

In this case, we have

$$\frac{n_2}{n_1} = \frac{\sigma_2^2}{\sigma_1^2} \approx \frac{1.203^2}{1.412^2} = 0.726.$$

So, if we originally needed $n = 100$, using the upper confidence bound reduces the sample size estimate to 72.6, rounded up to $n = 73$.

4. We are given samples of size $n_1 = 75$ and $n_2 = 45$ from independent normally distributed populations. Suppose we observe sample variances $S_1^2 = 412.09$ and $S_2^2 = 243.36$. Do a hypothesis test of

$$H_o : \sigma_2^2 = \sigma_1^2$$
$$H_a : \sigma_2^2 \neq \sigma_1^2$$

using an $\alpha = 0.05$ significance level. Give explicitly the rejection regions, and also report a P-value.

SOLUTION The test statistic is

$$F = \frac{S_1^2}{S_2^2} = \frac{412.09}{243.36} = 1.693,$$

which under H_o has a $F_{n_1-1,n_2-1} = F_{74,44}$ distribution. The relevant critical values for a two-sided size $\alpha = 0.05$ test are

$$\begin{aligned} F_{n_1-1,n_2-1,1-\alpha/2} &= F_{74,44,0.975} \approx 0.597 \\ F_{n_1-1,n_2-1,\alpha/2} &= F_{74,44,0.025} \approx 1.736. \end{aligned}$$

Since $F \geq F_{n_1-1,n_2-1,1-\alpha/2}$ and $F \leq F_{n_1-1,n_2-1,1-\alpha/2}$ we do not reject the null hypothesis at an α significance level. The P-value is 0.0611.

5. We are given samples of size $n_1 = 75$ and $n_2 = 103$ from independent normally distributed populations. Suppose we observe sample variances $S_1^2 = 299.29$ and $S_2^2 = 158.76$. Do a hypothesis test of

$$H_o : \sigma_2^2 = \sigma_1^2$$
$$H_a : \sigma_2^2 \neq \sigma_1^2$$

using an $\alpha = 0.05$ significance level. Give explicitly the rejection regions, and also report a P-value.

SOLUTION The test statistic is

$$F = \frac{S_1^2}{S_2^2} = \frac{299.29}{158.76} = 1.885,$$

which under H_o has a $F_{n_1-1,n_2-1} = F_{74,102}$ distribution. The relevant critical values for a two-sided size $\alpha = 0.05$ test are

$$\begin{aligned} F_{n_1-1,n_2-1,1-\alpha/2} &= F_{74,102,0.975} \approx 0.648 \\ F_{n_1-1,n_2-1,\alpha/2} &= F_{74,102,0.025} \approx 1.52. \end{aligned}$$

Since $F \geq F_{n_1-1,n_2-1,1-\alpha/2}$ we reject the null hypothesis at an α significance level. The P-value is 0.00304.

6. Write an `R` function that accepts two samples as arguments, and returns the respective sample variances and sample sizes, the F statistic appropriate for an equality of variance test, and the p-value for a two-sided test against null hypothesis $H_o : \sigma_1^2 = \sigma_2^2$. Test your function using the following two samples.

```
x = c(6.00, 6.95,8.72,10.83)
y = c(19.38, 19.39, 20.04, 20.57, 19.91, 20.05, 19.82, 20.20)
```

SOLUTION The following function implements the required F test

```
ftest = function(x,y) {

  n1 = length(x)
  n2 = length(x)

  var1 = var(x)
  var2 = var(y)

  fstat = var1/var2
  pval = 2*min(pf(fstat,n1-1,n2-1),1-pf(fstat,n1-1,n2-1))
  return(c(var1=var1, var2=var2, n1=n1, n2=n2, fstat=fstat, pval=pval))
}
```

The following code gives the required example:

```
> x = c(6.00, 6.95,8.72,10.83)
> y = c(19.38, 19.39, 20.04, 20.57, 19.91, 20.05, 19.82, 20.20)
> ftest(x,y)
       var1        var2          n1          n2       fstat        pval
 4.52243333  0.15925714  4.00000000  4.00000000 28.39705179  0.02108341
>
```

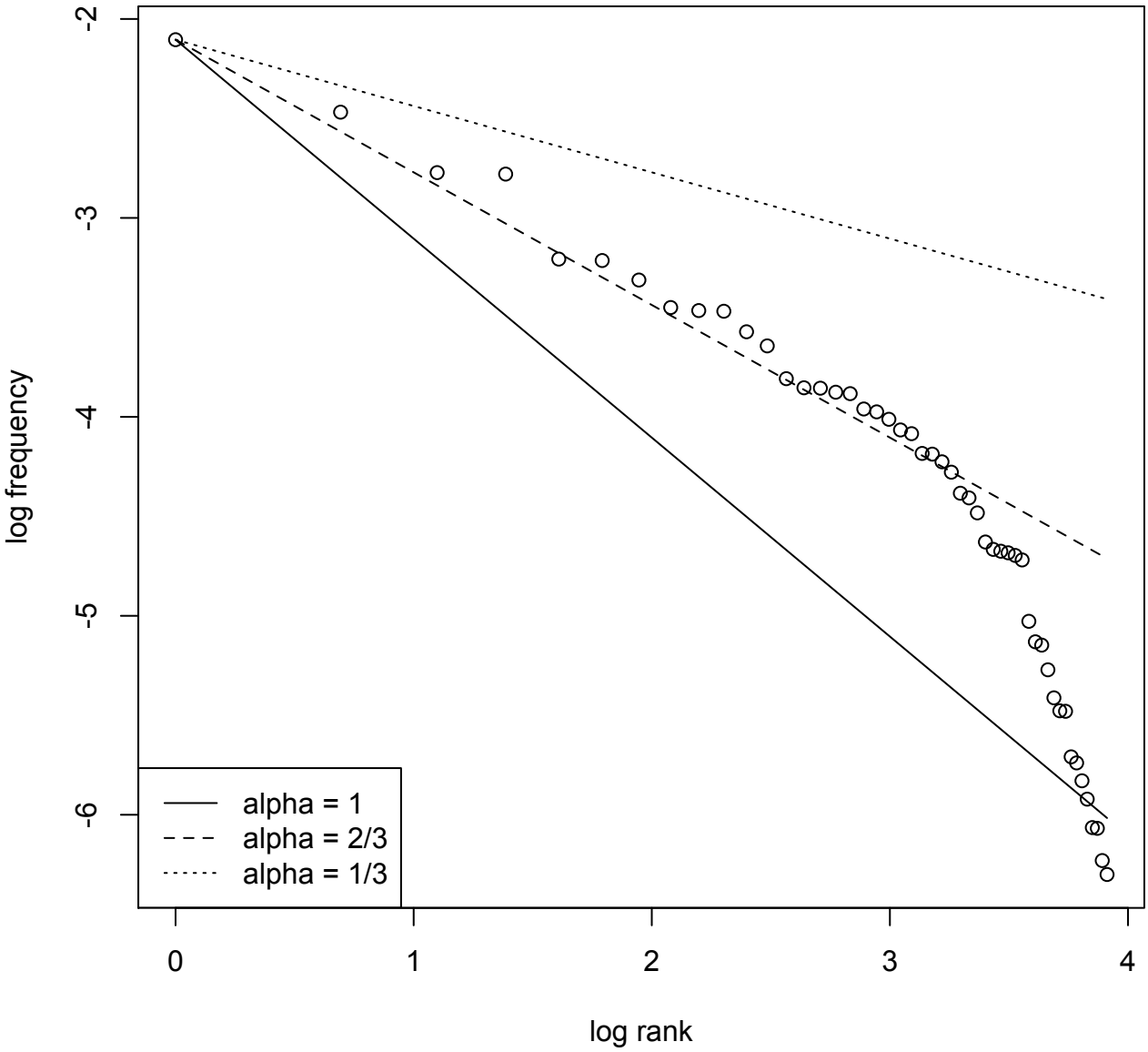


Figure 1: Plot for Problem 6.

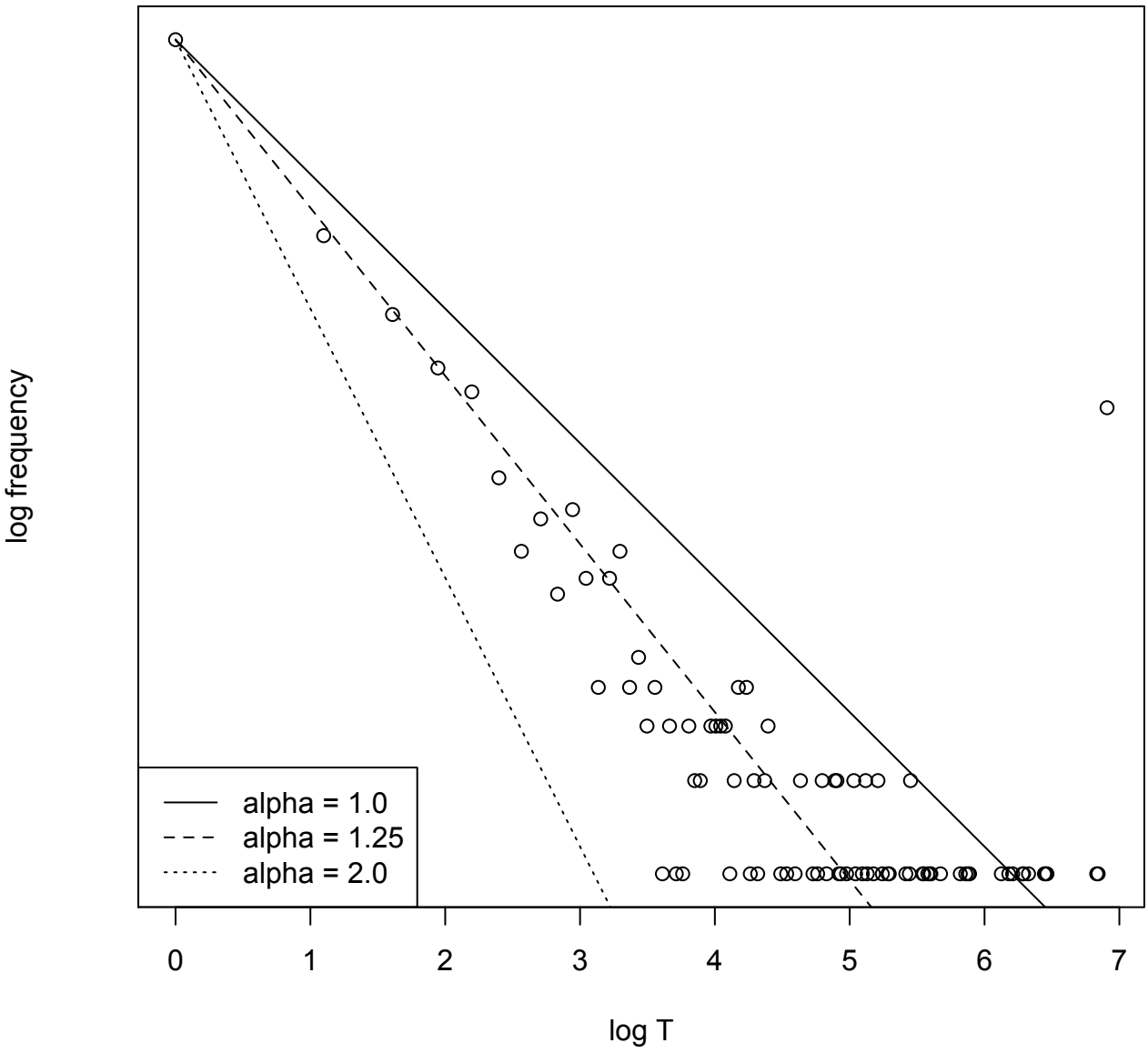
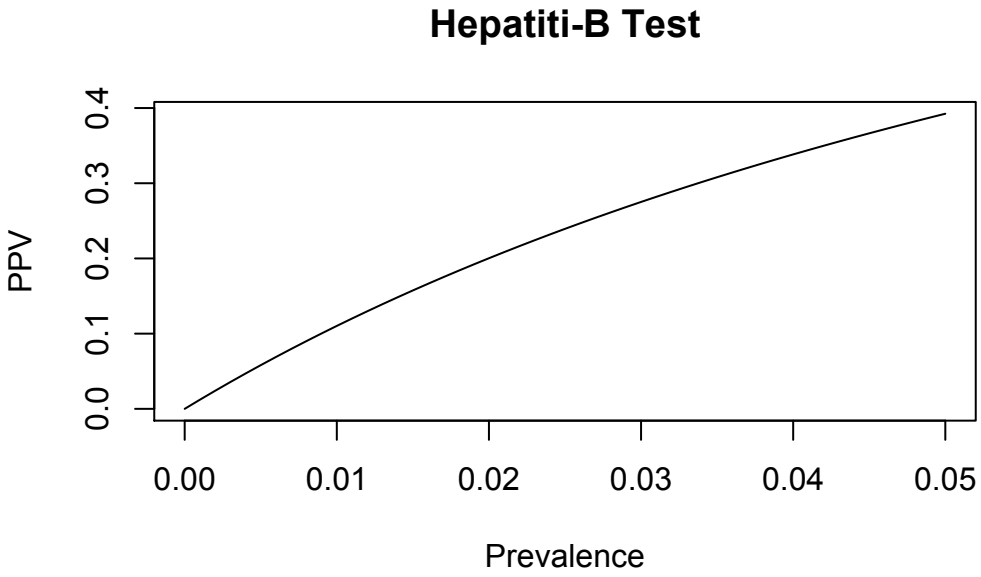


Figure 2: Plot for Problem 7.



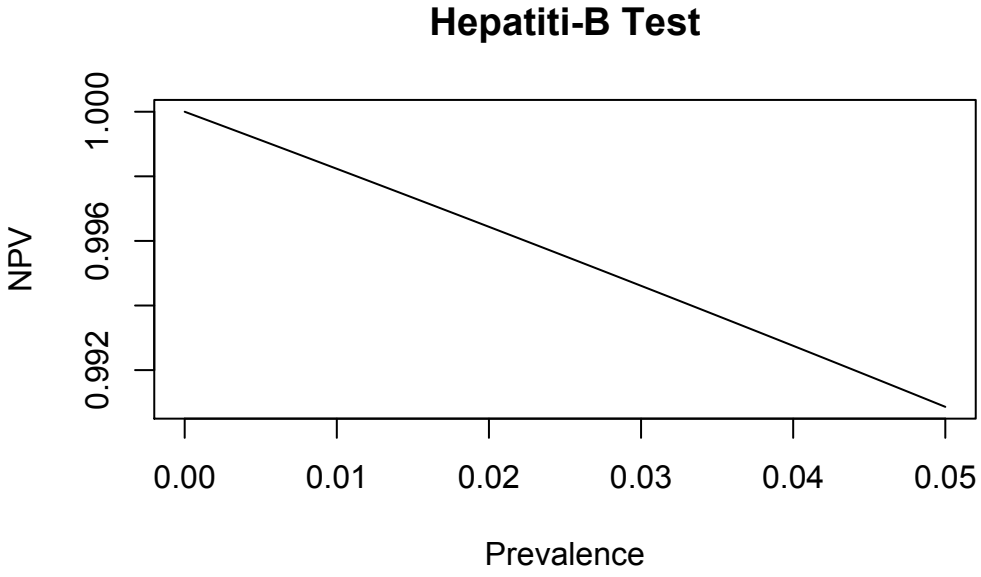


Figure 3: Plot for Problem 8.

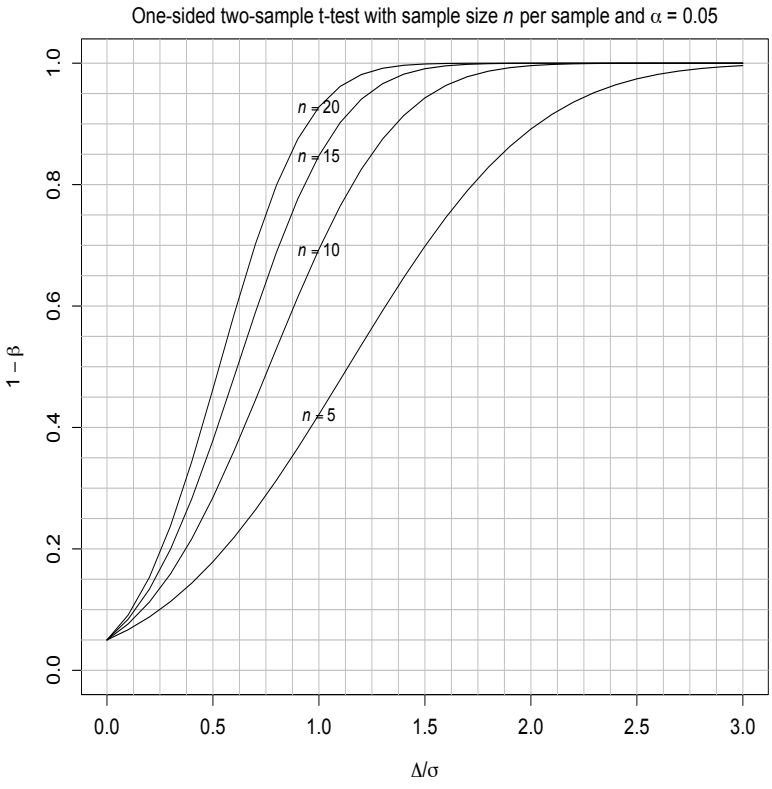


Figure 4: Plot for Problem 1.

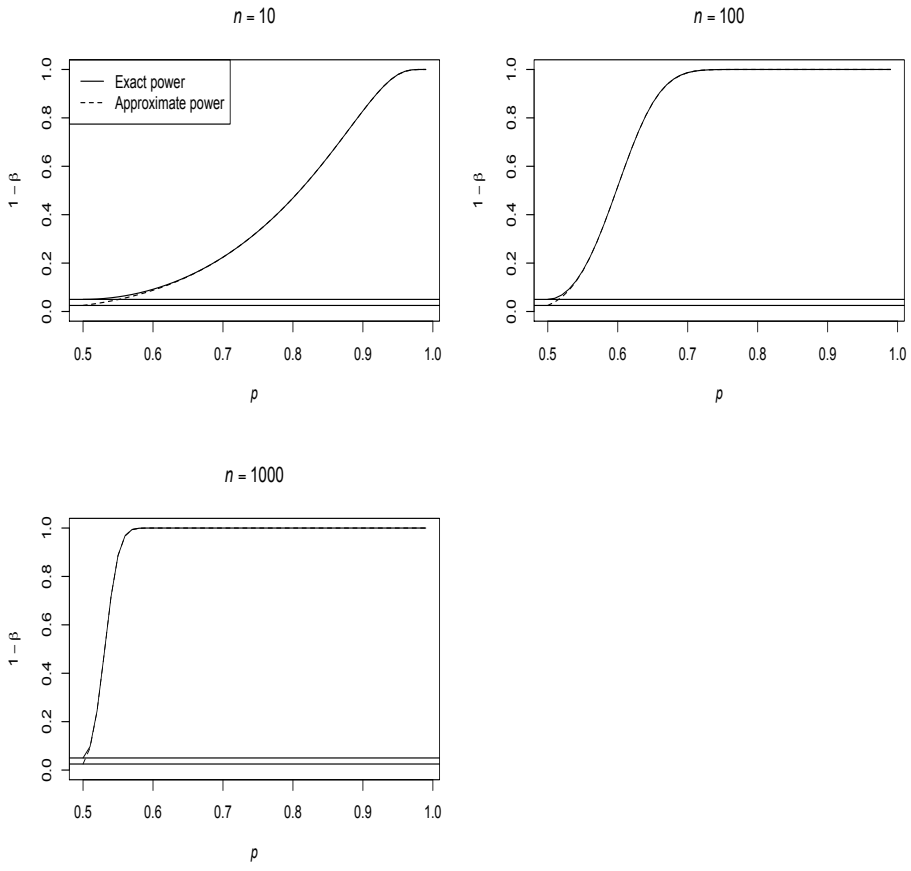


Figure 5: Plot for Problem 2.